a S.A.F.E. approach to risk: Saddlepoint Approximations in Financial Econometrics

Dissertation
der Wirtschaftswissenschaftlichen Fakultät
der Universität Zürich

zur Erlangung der Würde
eines Doktors der Ökonomie

vorgelegt von
Simon A. Broda
aus Deutschland

Genehmigt auf Antrag von
Prof. Dr. Marc Paolella
Prof. M. Wolf, Ph. D.

Die Wirtschaftswissenschaftliche Fakultät der Universität Zürich gestattet hierdurch die Drucklegung der vorliegenden Dissertation, ohne damit zu den darin ausgesprochenen Anschauungen Stellung zu nehmen.

Zürich, den 22. Oktober 2007

Der Dekan: Prof. Dr. H. P. Wehrli

To the Memory of my Father

Acknowledgements

The preparation of a thesis is a task one could not hope to master without support, and mine is certainly no exception. Fortunately, I have enjoyed the best support that one could wish for. I am deeply indebted to my supervisor Marc Paolella, for introducing me to the topic of saddlepoint approximations, for giving me the opportunity to work on this thesis, and for his guidance in my research, and to my co-supervisor, Michael Wolf, for his many helpful comments, in particular on the second chapter of this thesis. My colleagues Yianna Tchopourian, Ramona Westermann, Matteo Bonato, and Sven Steude, deserve my gratitude not only for the fruitful discussions that we shared, but, above all, for being the great friends that they are. Finally, I would like to thank my parents, Peter and Elvira Broda, my sister Mirjam, and my wife Tzvetelina, for the love and support they bestowed on me, and for bearing with me through the ups and downs of writing this dissertation.

Without you, this thesis could not have been written.

Contents

Introduction

Beginning with the seminal paper of Daniels (1954), and with an increasing number of publications in the final 20 years of the past millennium, saddlepoint approximations have become a valuable tool in statistics. However, despite more than five decades of active development, their use in financial econometrics is still far from widespread, with the few notable exceptions mostly confined to the field of credit risk (see, e.g. Gordy, 2002; Glasserman, 2004). The aim of this collection of manuscripts is to demonstrate that these approximations can be of great use in other fields within this discipline, in areas as far apart as empirical corporate finance and Value at Risk forecasting, and with applications ranging from approximating critical values of tests, to modelling stock returns.

The technicalities of the approximation are deferred to subsequent chapters; suffice it to say here that saddlepoint approximations offer a viable alternative whenever the exact density or distribution function of a random variable of interest is either intractable, or computationally too demanding. This occurs frequently when dealing with convolutions of independent and identically distributed random variables, or when a probability distribution is defined in terms of certain underlying random variables, as in the case of the noncentral Student's t. In such cases, the saddlepoint method offers a way of approximating the desired distribution, requiring only the existence (and tractability) of the moment generating function of the random variables involved. The advantage of the saddlepoint approximation over alternative approximations is that it has *relative* error, as opposed to the *absolute* error pertinent to, e.g., the Edgeworth expansion, which therefore tends to give poor results in the tails of the distribution. Clearly, in financial applications, interest typically centers around the extreme events embodied in those very tails, making the saddlepoint approximation a natural tool in such endeavors.

The remainder of this thesis is divided into two parts, each of which contains two separate manuscripts. Manuscripts 1 and 2, which are reproduced in Part I, deal with autoregressive models, in the pure time-series and panel settings, respectively. The methods developed in the latter, especially, have applications in empirical corporate finance, for example in testing the trade-off versus the pecking order theory of capital structure (see, e.g., Shyam-Sunder and Myers, 1999).

In particular, Manuscript 1 develops a new point estimator for the AR(1) coefficient in the linear regression model with arbitrary exogenous regressors and stationary AR(1) disturbances. It

is well known that the standard (least squares) estimator of this parameter exhibits a pronounced downward bias. The new estimator is designed to alleviate this problem; its construction parallels that of the median-unbiased estimator of Andrews (1993), but uses the mode as a measure of central tendency. A mean-adjusted estimator is also considered, and saddlepoint approximations are used to lower the computational burden of all the estimators. This facilitates a large-scale simulation study for assessing the small-sample properties of the competing estimators. It is demonstrated that their relative performance depends almost exclusively on the value of the autoregressive parameter, with the new estimator dominating over a large part of the parameter space. A shortened version of this manuscript has appeared in the Journal of Computational Statistics and Data Analysis (Volume 51, Issue 7, 2007, Pages 3355–3367, with M. Paolella and K. Carstensen).

In Manuscript 2, these methods are extended to the panel case, i.e., to models with both a cross-sectional and a time-series dimension, where the problem of bias is even more acute. Based on the notion of a quantile unbiased estimating function, a general method is developed for conducting exact small-sample inference in models which allow the estimator of the (scalar) parameter of interest to be expressed as the root of an estimating function, and which is particularly simple to implement for linear models with a covariance matrix depending on a single parameter. The method, dubbed quantile unbiased estimation, or QUEST, involves the computation of tail probabilities of the estimating function. In the context of dynamic panel models, both the least squares and maximum likelihood paradigms give rise to estimating functions involving sums of ratios in quadratic forms in normal variates, the distribution of which cannot be straightforwardly computed. This obstacle is overcome by a saddlepoint approximation that is both readily evaluated and remarkably accurate. A simulation study demonstrates the validity of the procedure.

Part II is dedicated to generalized autoregressive conditional heteroskedasticity, or GARCH, models, which can be considered the workhorse of financial time series analysis. Specifically, Manuscript 3 is dedicated to univariate models. As is well known, financial returns data sampled at weekly or higher frequency tend to exhibit excess kurtosis and moderate skewness, thus violating the assumption of Gaussianity, and hence prompting the adoption of different, non-Gaussian innovations distributions able to accommodate these features. One such distribution is the doubly noncentral t, but its evaluation is computationally costly. Based on saddlepoint methods, this manuscript derives closed-form approximations for both its density and cumulative distribution function. They exhibit remarkable accuracy throughout the entire support of the distribution, and are vastly superior to existing approximations. The enormous increase in computational speed facilitates use of this distribution in the maximum likelihood estimation of GARCH-type models, as is pursued here for a data set of daily NASDAQ returns. A shortened version of this manuscript has appeared in the Journal of Computational Statistics and Data Analysis (Volume 51, Issue 6, 2007, Pages 2907–2918, with M. Paolella).

Above method is, however, constrained to univariate GARCH models. The estimation of multivariate GARCH models is a considerably more challenging task, even in modern computer environments, and requires a custom solution. One such solution is considered in Manuscript 4. It is shown how Independent Component Analysis can be used to estimate the Generalized Orthogonal GARCH model in a fraction of the time otherwise required. The proposed method is a two-step procedure, separating the estimation of the correlation structure from that of the univariate dynamics, and thus facilitating the incorporation of non-Gaussian innovations distributions in a simple and efficient manner. The generalized hyperbolic distribution provides an excellent parametric description of financial returns data and is used for the univariate fits, but its convolutions, necessary for portfolio risk calculations, are intractable. In order to overcome this restriction, a saddlepoint approximation to the required distribution function is derived, which is both computationally cheap and extremely accurate — most notably in the tail, which is crucial for risk calculations. A simulation study and an application to stock returns data corroborate the usefulness of the method.

Part I

Autoregressive Models

Manuscript 1

Bias–Adjusted Estimation in the ARX(1) Model

1.1 Introduction

Inferential procedures for the parameters of autoregressive models (with or without covariates) continue to receive a great amount of attention in the theoretical literature, with recent contributions including work on structural breaks (Breitung, 2002; Kurozumi, 2002; Saikkonen and Lütkepohl, 2002 and the references therein), innovation variance shifts (Kim et al., 2002), methods which use point optimal tests (Shively, 2001), local-to-unity arguments (Elliot and Stock, 2001), and bootstrap techniques (Hansen, 1999).

This paper develops a new estimator for the autoregressive coefficient in a first–order autoregression. Its derivation parallels that of the median–unbiased point estimator of Andrews (1993), but uses the mode, rather than the median, as a measure of central tendency. Both of these methods entail the relatively costly numeric evaluation of the distribution function of a ratio of quadratic forms in normal random variables, accomplished by inversion of the relevant characteristic function via, for example, the method of Imhof (1961). A way of circumventing these extensive calculations — without the restrictions associated with pre–computed tables, as provided by Andrews for his estimator — is to replace the exact evaluation of the requisite distribution function with a saddlepoint approximation, thus removing the bottleneck in the procedure so that the estimators can be calculated in about a hundredth of the time otherwise necessary. Along with these massive time savings, the accuracy of the saddlepoint approximation, or, in short, SPA, is not only high enough for practical work, but in fact results in *higher* accuracy than can be achieved by interpolation from pre–computed tables.

Because of the similarity in its construction, we also include in our study the mean–adjusted estimator of Tanizaki (2000), and show how it can be computed without having to resort to simulation. Owing to these numerical methods, all three bias corrected estimators can be computed fast enough to make a simulation study feasible, using a variety of data generating models. We find, somewhat surprisingly—and quite conveniently—, that the relative small–sample properties of the estimators is *virtually invariant to the choice of sample size and set of regressors*. Moreover, this fortuitous behavior remains (approximately) constant for a variety of non-Gaussian innovation distribution assumptions commonly entertained in practice. The optimal choice of estimator depends (essentially) only on the true value of the autoregressive parameter α, *but in virtually the same way for any model design and distributional assumption*. For example, one of the estimators has lowest mean squared error for all α between about 0.7 and 1.0—a result which should be of interest when working with series with high persistence or near unit–root behavior. The ranges of α for which a particular estimator is optimal are all quite large, these being $(-1, -0.1)$, $(-0.1, 0.7)$ and $(0.7, 1)$, so that even a very small amount of "prior information" on the part of the researcher can be effectively used.

The remainder of this paper is as follows. Section 1.2 introduces the model and relevant notation. Section 1.3 and 1.4 review the median–unbiased and mean–adjusted estimators previously proposed in the literature, respectively, and show how the computational burden associated with

them can be substantially reduced. Section 1.5 develops the mode–adjusted estimator. Section 1.6 details several simulation studies, from which the performance of the estimators can be assessed and recommendations made for their use. Section 1.7 concludes and briefly discusses some ideas for further research. Three short appendices provide (i) the proof of an invariance property of the least squares estimator for α, (ii) formulae for the saddlepoint approximations and (iii) computational details on the simulation studies.

1.2 The Model

Using notation similar to that in Andrews (1993), the model consists of an observed and a latent equation given respectively by

$$Y_t = \mathbf{x}_t'\boldsymbol{\beta} + Y_t^\ell, \qquad t = 0, \ldots, T, \tag{1.1}$$

and

$$Y_t^\ell = \alpha Y_{t-1}^\ell + U_t, \qquad t = 1, \ldots, T, \qquad U_t \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2\right), \tag{1.2}$$

where $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_T]'$ is assumed to be a full rank $(T+1) \times k$ matrix, and the initial observation Y_0^ℓ follows the unconditional distribution of Y_t, $Y_0^\ell \sim \mathrm{N}\left(0, \sigma^2/\left(1-\alpha^2\right)\right)$ if $\alpha \in (-1, 1)$ and an arbitrary constant if $\alpha = 1$. Note that a unit root $\alpha = 1$ implies that (1.1) becomes a spurious regression model. Asymptotically, in such a model, the effect of x_t on y_t is irrelevant and the parameter $\boldsymbol{\beta}$ is not meaningful to interpret. However, we work with finite samples, and our main focus is on inference for α. Moreover, while we do not engage in unit root testing in this paper, it is straightforward to extend the methods described herein to the construction of confidence intervals, and thus, tests of the unit root hypothesis; see Andrews (1993).

We consider point estimators for α. Point- and interval estimates for the regression coefficients $\boldsymbol{\beta}$ can be obtained by generalized least squares, using the estimated value of α in the requisite covariance matrix. However, in our experiments, the differences in inferential accuracy when using different estimators for α were negligible; as such, we focus our attention on the autoregressive parameter.

The AR(1) model (1.1)–(1.2) can be estimated by least squares after combining the observable and latent equations to

$$Y_t = Y_{t-1}\alpha + \mathbf{x}_t'\boldsymbol{\beta} - \mathbf{x}_{t-1}'\boldsymbol{\beta}\alpha + U_t, \quad t = 1, \ldots, T,$$

or, in matrix form,

$$\mathbf{Y}_T = \mathbf{Y}_{T-1}\alpha + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{U}_T, \tag{1.3}$$

where $\gamma = [\boldsymbol{\beta}', -\boldsymbol{\beta}'\alpha]'$, $\mathbf{Y}_T = [Y_1, \ldots, Y_T]'$, $\mathbf{Y}_{T-1} = [Y_0, \ldots, Y_{T-1}]'$, $\mathbf{Z} = [\mathbf{X}_T, \mathbf{X}_{T-1}]$, $\mathbf{X}_T = [\mathbf{x}_1, \ldots, \mathbf{x}_T]'$, $\mathbf{X}_{T-1} = [\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}]'$ and $\mathbf{U}_T = [U_1, \ldots, U_T]'$.

It should be noted that, due to the common factor restrictions on the $2k$ parameters in $\boldsymbol{\gamma}$, model (1.3) is, in general, different from the dynamic linear model

$$Y_t = Y_{t-1}\alpha + \mathbf{x}_t'\boldsymbol{\beta} + U_t, \quad t = 1, \ldots, T. \tag{1.4}$$

Neglecting these restrictions, a straightforward application of the Frisch–Waugh theorem as in Andrews (1993) shows that the ordinary least squares (OLS) estimator of α can be expressed as

$$\hat{\alpha}_{\mathrm{LS}} = \frac{\mathbf{Y}_{T-1}'\mathbf{M}\mathbf{Y}_T}{\mathbf{Y}_{T-1}'\mathbf{M}\mathbf{Y}_{T-1}}, \tag{1.5}$$

where $\mathbf{M} = \mathbf{I}_T - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$. If $\mathbf{Z}$ has reduced column rank, then it should be replaced by a full column rank matrix $\tilde{\mathbf{Z}}$ spanning the same column space. As remarked by Andrews (1993, p. 146), and as is required for construction of the estimators considered herein, $\hat{\alpha}_{\mathrm{LS}}$ is independent of $\boldsymbol{\beta}$ and σ^2 (and of Y_0 if $\alpha = 1$) for any exogenous regressor matrix $\mathbf{X}$. A detailed proof is given in Appendix 1.A. Andrews (1993, p. 146) also notes that the methodology can be made applicable to model (1.4) by including in $\mathbf{Z}$ a set of constructed regressors, designed to maintain the invariance property. For the specification of these regressors, see Kiviet and Phillips (1990).

1.3 Median–unbiased Estimation

It is well known that $\hat{\alpha}_{\mathrm{LS}}$ is downward biased, extremely so for α near one. While various procedures exist to partially correct for this, no operational method has so far been devised which is exactly mean–unbiased. It is, however, straightforward to construct a median–unbiased estimator, hereafter denoted $\hat{\alpha}_{\mathrm{Med}}$, first pursued in this context by Andrews (1993). By definition, an estimator $\hat{\theta}$ is median–unbiased for θ if, for each value θ in the parameter space, θ is a median of $\hat{\theta}$. The following bias correction procedure then makes $\hat{\alpha}_{\mathrm{Med}}$ a median–unbiased estimator: $\hat{\alpha}_{\mathrm{Med}}$ takes that value of α that yields the OLS estimator to have a median equal to the OLS estimate obtained from the data. More formally, let $\mathrm{Med}\left(\hat{\alpha}_{\mathrm{LS}} \mid \alpha, \mathbf{X}\right) = m\left(\alpha\right)$ denote the median function of $\hat{\alpha}_{\mathrm{LS}}$ when α is the true parameter, and let $m^{-1} : \left(m\left(-1\right), m\left(1\right)\right] \to \left(-1, 1\right]$ denote its inverse. For this to be meaningful, it is required that $m\left(\alpha\right)$ be strictly increasing. As noted by Andrews, it is not apparent how the latter condition can be verified analytically; however, in the present setup, numerical evidence suggests it holds.

The median unbiased estimator $\hat{\alpha}_{\mathrm{Med}}$ is then given by

$$\hat{\alpha}_{\mathrm{Med}} = \begin{cases} 1, & \text{if } \hat{\alpha}_{\mathrm{LS}} > m\left(1\right), \\ m^{-1}\left(\hat{\alpha}_{\mathrm{LS}}\right), & \text{if } m\left(-1\right) < \hat{\alpha}_{\mathrm{LS}} \leq m\left(1\right), \\ -1, & \text{if } \hat{\alpha}_{\mathrm{LS}} \leq m\left(-1\right). \end{cases} \tag{1.6}$$

Given the observed value of the OLS estimator, say $\hat{\alpha}_{\mathrm{LS}}^{O}$, the estimator can be expressed for

$m(-1) < \hat{\alpha}_{\mathrm{LS}}^{O} \leq m(1)$ as

$$\hat{\alpha}_{\mathrm{Med}} = m^{-1}(\hat{\alpha}_{\mathrm{LS}}) = \mathrm{argmin}_{\alpha} \left| \mathrm{Med}(\hat{\alpha}_{\mathrm{LS}} \mid \alpha, \mathbf{X}) - \hat{\alpha}_{\mathrm{LS}}^{O} \right|. \tag{1.7}$$

Equivalently, with $F_{\hat{\alpha}_{\mathrm{LS}}}$ denoting the cumulative distribution function (cdf) of $\hat{\alpha}_{\mathrm{LS}}$,

$$m^{-1}(\hat{\alpha}_{\mathrm{LS}}) = \mathrm{argmin}_{\alpha} \left| F_{\hat{\alpha}_{\mathrm{LS}}}\left(\hat{\alpha}_{\mathrm{LS}}^{O} \mid \alpha, \mathbf{X}\right) - 0.5 \right|, \tag{1.8}$$

which is more suitable for computation than (1.7).

To get an idea of the magnitude of the correction induced by $\hat{\alpha}_{\mathrm{Med}}$, the top panel of Figure 1.1 plots values of $\hat{\alpha}_{\mathrm{LS}}$ on the ordinate (y-axis) versus the corresponding quantity which should be added to $\hat{\alpha}_{\mathrm{LS}}$ to arrive at $\hat{\alpha}_{\mathrm{Med}}$ on the x-axis (note that the requirement that the median function be monotonic does not imply that the correction, as a function of the observed OLS estimator, is monotonic). For example, with $T = 10$ and an $\mathbf{X}$ matrix consisting of intercept and trend, if $\hat{\alpha}_{\mathrm{LS}} = 0.2$, then $\hat{\alpha}_{\mathrm{Med}} \approx 0.68$. As expected, the amount of correction decreases as the sample size increases. One also sees that, particularly for smaller sample sizes, the amount of correction vastly increases when the model changes from intercept to intercept and trend.

Evaluation of (1.8) is possible and straightforward if $F_{\hat{\alpha}_{\mathrm{LS}}}$ is computable. From expression (1.14) in Appendix 1.A, this involves the cdf of a ratio of quadratic forms in normal variables, which can be evaluated by numerically inverting an associated characteristic function, as detailed by Imhof (1961) in this context. As Andrews (1993) noted, such a computation is prohibitively slow, with simulation being a viable alternative. Because the three special regressor cases (no intercept, intercept, and intercept and time trend) arise frequently in applications, Andrews (1993) tabulated the necessary quantiles, to three significant digits, for a grid of ten (unequally spaced) sample sizes between 40 and 200 and 20 (unequally spaced) α values between -0.999 and 1.0. While two-dimensional interpolation of the tabulated values can easily be automated in a computer, it will still result in only about two digit accuracy. Moreover, for sample sizes outside the range $[40, 200]$ or—more likely—a different set of regressors, the tables are not applicable.

To address these shortcomings, use can be made of the so–called saddlepoint approximation for evaluating $F_{\hat{\alpha}_{\mathrm{LS}}}$. The saddlepoint method can be viewed as an accurate approximation to the inversion of the characteristic function, but without the need for numerical integration, which gives rise to its enormous speed advantage (see Barndorff-Nielsen and Cox, 1989, Jensen, 1995, and Goutis and Casella, 1999, for a general overview, and Lieberman, 1994b, for the SPA in the context we use herein). It does, however, require the existence of the moment generating function, which holds in the setting considered in this paper, but not, in general, for all random variables. An understanding of the SPA is not necessary to use our proposed methods; the appendix contains the required formulae for the computations and relevant references to the literature. Programs (in Matlab) are also available from the authors to perform all the calculations.

As mentioned, the SPA is not exact. It yields two to three digit accuracy for sample sizes

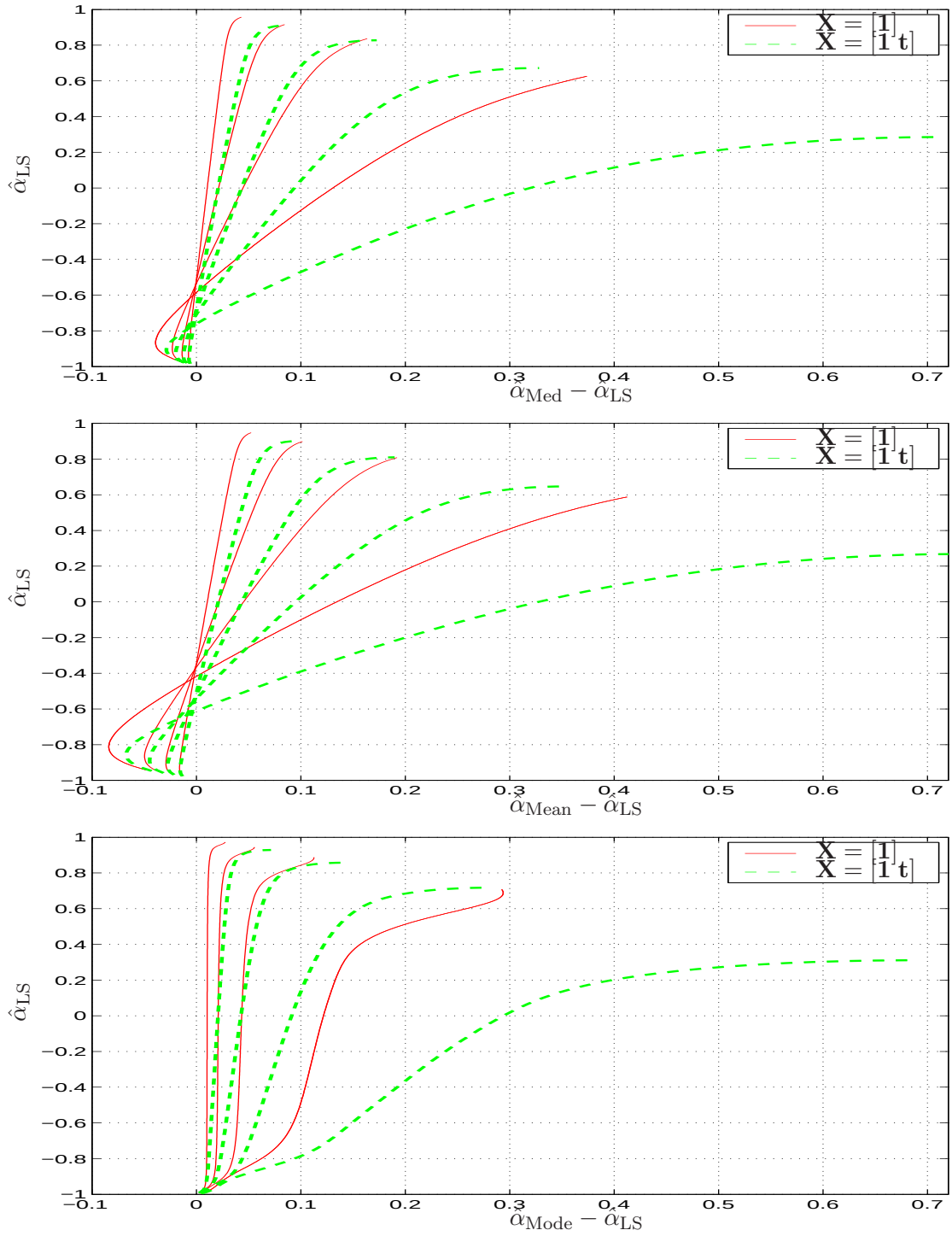

Figure 1.1: Adjustment to $\hat{\alpha}_{\mathrm{LS}}$ corresponding to $\hat{\alpha}_{\mathrm{Med}}$ (top), $\hat{\alpha}_{\mathrm{Mean}}$ (middle) and $\hat{\alpha}_{\mathrm{Mode}}$ (bottom), shown for the two $\mathbf{X}$ matrices constant (solid lines) and constant/trend (dashed lines). The four sample sizes shown are $T = 10$, $T = 25$, $T = 50$ and $T = 100$, moving from right to left. See also the text in Section 1.3 for explanation.

between 10 and 30, and from three to four digit accuracy for samples of between 50 and 80 observations (and is asymptotically exact), and so matches (or usually exceeds) that obtained from using linear interpolation from the tables in Andrews (1993). Depending on the method used for numerical integration and the specified tolerance on the error in the Imhof (1961) routine, the SPA is between 10 and 200 times faster, and is easier to program, requiring only a univariate root search and evaluation of the standard normal cdf, both of which are implemented in virtually all statistical computing packages.

1.4 Mean–adjusted Estimation

For inference on α via the statistic $\hat{\alpha}_{\rm LS}$, it appears infeasible to construct a mean–unbiased estimator, but a procedure which comes very close (and turns out to exhibit other good small–sample properties) has been proposed by Tanizaki (2000), and also by MacKinnon and Smith (1998), in a more general context. It amounts to interpreting $m(\cdot)$ as the analogously defined mean function in (1.6), i.e., let $m(\alpha) = \mathbb{E}[\hat{\alpha}_{\rm LS} \mid \alpha, \mathbf{X}]$. Like the median function, numerical results suggest that it is strictly increasing for $-1 < \alpha < 1$, so that its inverse exists. In particular, for $m(-1) < \hat{\alpha}_{\rm LS}^{O} \leq m(1)$,

$$\hat{\alpha}_{\rm Mean} = m^{-1}(\hat{\alpha}_{\rm LS}) = \operatorname{argmin}_{\alpha} \left| \mathbb{E}[\hat{\alpha}_{\rm LS} \mid \alpha, \mathbf{X}] - \hat{\alpha}_{\rm LS}^{O} \right|, \tag{1.9}$$

which we refer to as the *mean–adjusted* estimator. It is not exactly mean–unbiased because of the truncation at -1 and 1 and because of the nonlinearity of the mean function, i.e., $\mathbb{E}\left[m^{-1}(\hat{\alpha}_{\rm LS})\right] \neq m^{-1}(\mathbb{E}[\hat{\alpha}_{\rm LS}]) = \alpha$.

Tanizaki used simulation to obtain the mean function $\mathbb{E}[\hat{\alpha}_{\rm LS}]$ in (1.9). Due to sampling variation, the inversion of the mean function obtained in that way is prone to instability, rendering a faster and more reliable method for its evaluation desirable. As such, we suggest to use the expressions for the first and second moments of a ratio of central quadratic forms in normal variables as given in Sawa (1978). Specifically, with $\mathbf{A}$ and $\mathbf{B}$ as defined in (1.14) in Appendix 1.A, let $\mathbf{P}\mathbf{\Lambda}\mathbf{P}'$ be the spectral decomposition of $\mathbf{B}$ and set $\mathbf{C} = \mathbf{P}'\mathbf{A}\mathbf{P}$. Then

$$\mathbb{E}[\hat{\alpha}_{\rm LS}] = \int_0^{\infty} \sum\nolimits_{j=1}^{T+1} \frac{c_j}{(1+2\lambda_j t)^{3/2} \prod\limits_{k \neq j} (1+2\lambda_k t)^{1/2}} dt, \tag{1.10}$$

where c_j and λ_j denote the $j^{\rm th}$ diagonal element of $\mathbf{C}$ and $\mathbf{\Lambda}$, respectively. The indefinite integral in (1.10) can be evaluated directly, by using the fact that most of the integrand mass is near zero and that the integrand dies off rapidly, or by transforming the range of t to lie in an open interval of finite length (e.g., via the substitution $u = 1/(1+t)$). The former approach was found to be faster and numerically more reliable; see also Paolella (2003).

The middle panel of Figure 1.1 is similar to the top panel, but shows the correction appropriate

for $\hat{\alpha}_{\text{Mean}}$. While certainly different, it differs significantly from the top panel only for values of $\hat{\alpha}_{\text{LS}}$ less than -0.4.

1.5 Mode–adjusted Estimation

Use of bias adjustment methods based on the mean and median (as measures of central tendency) leads naturally to consideration of the third such measure: the mode. Following (1.7) and (1.9), it is natural to define the *mode–adjusted* estimator as

$$\hat{\alpha}_{\text{Mode}} = m^{-1}\left(\hat{\alpha}_{\text{LS}}\right) = \text{argmin}_{\alpha} \left|\text{Mode}\left(\hat{\alpha}_{\text{LS}} \mid \alpha, \mathbf{X}\right) - \hat{\alpha}_{\text{LS}}^{O}\right|, \tag{1.11}$$

where $m(\cdot)$ is now interpreted in (1.6) as the mode function. To the best of the authors' knowledge, such an estimator has not been previously proposed. In comparison to $\hat{\alpha}_{\text{Med}}$ and $\hat{\alpha}_{\text{Mean}}$, which are well–defined and unique for continuous distributions with finite first moment, use of $\hat{\alpha}_{\text{Mode}}$ only makes sense if the relevant distribution is unimodal. Indeed, inspection shows that, for sample sizes greater than five, the probability density function (pdf) of $\hat{\alpha}_{\text{LS}}$ is unimodal and, paralleling the requirements for the median- and mean-adjusted estimators, the mode function of $\hat{\alpha}_{\text{LS}}$ is strictly increasing for $|\alpha| < 1$, thus guaranteeing that $\hat{\alpha}_{\text{Mode}}$ is uniquely defined.

Let $f_{\hat{\alpha}_{\text{LS}}}(x;\alpha)$ denote the pdf of $\hat{\alpha}_{\text{LS}}$ at x when the true parameter is α (and suppress the dependency on the $\mathbf{X}$–matrix). From the definition of the mode, it follows that (1.11) is equivalent to choosing $\hat{\alpha}$ such that the density $f_{\hat{\alpha}_{\text{LS}}}(x;\hat{\alpha})$ attains its maximum at the observed value of $\hat{\alpha}_{\text{LS}}$. That is, we can write

$$\hat{\alpha}_{\text{LS}}^{O} = \text{argmax}_{x}\, f_{\hat{\alpha}_{\text{LS}}}(x;\hat{\alpha}_{\text{Mode}}), \tag{1.12}$$

i.e., $\hat{\alpha}_{\text{Mode}}$ is the (unique) value of α such that the observed value is a mode of $f_{\hat{\alpha}_{\text{LS}}}(x;\alpha)$.

Under the stated assumptions of unimodality and monotonicity of the mode of $\hat{\alpha}_{\text{LS}}$ as a function of α (for $|\alpha| < 1$), $\hat{\alpha}_{\text{Mode}}$ is the unique solution to the implicit equation

$$\left.\frac{\partial f_{\hat{\alpha}_{\text{LS}}}(x;\hat{\alpha}_{\text{Mode}})}{\partial x}\right|_{x=\hat{\alpha}_{\text{LS}}^{O}} = 0, \tag{1.13}$$

which can be solved by a univariate root search in much the same fashion as is required for $\hat{\alpha}_{\text{Med}}$ and $\hat{\alpha}_{\text{Mean}}$.

Solving equations (1.12) or (1.13) is not trivial, as closed–form expressions for Mode$(\hat{\alpha}_{\text{LS}})$ or $f_{\hat{\alpha}_{\text{LS}}}(x;\alpha)$ do not exist. Regarding the latter, because the characteristic function of a ratio of quadratic forms is numerically intractable, standard inversion formulae for evaluating the pdf cannot be applied, as is possible for the cdf. We circumvent this problem by again using a saddlepoint approximation: A first and second order SPA to the pdf of a ratio of quadratic forms in zero–mean normal random variables has been constructed by Lieberman (1994b). Appendix 1.B contains all the relevant expressions.

The one potential drawback of using the SPA for the pdf is that neither the first nor second order expression integrates precisely to one, although both are usually very close (the latter even more so), but the exact integrating constant depends on the true value of α and on the $\mathbf{X}$ matrix, and so cannot be determined without numerical integration for each case. Fortunately, for computing $\hat{\alpha}_{\text{Mode}}$, this issue is irrelevant, because the position of the maximum of the density will not change with normalization. Furthermore, the SPA to the pdf is continuous over the (interior of the) entire support (unlike the cdf, which requires some finessing near the mean; see Appendix 1.B).

Thus, computation of $\hat{\alpha}_{\text{Mode}}$ can be fully operationalized in this context in a numerically fast and accurate fashion. The only possible remaining caveat to its effective use is the approximate nature of the density via the SPA. To check this, we used the numerical second derivative of the exact cdf of $\hat{\alpha}_{\text{LS}}$, which can be made numerically reliable enough for approximating the mode, but is extremely time consuming compared to use of the SPA. Using the very small sample size of $T = 25$, we found that differences in $\hat{\alpha}_{\text{Mode}}$ based on the SPA and use of the exact cdf occurred only in the third to fourth decimal place, thus confirming that use of the SPA in this context will not jeopardize the accuracy of the method by any appreciable amount. (As T increases, so does the accuracy of the SPA, because the distribution of $\hat{\alpha}_{\text{LS}}$ approaches the normal, for which the SPA is exact.)

The bottom panel of Figure 1.1 shows the correction appropriate for $\hat{\alpha}_{\text{Mode}}$. Notice how it differs considerably from the other two, implying that its small–sample properties should also differ markedly from those of $\hat{\alpha}_{\text{Med}}$ and $\hat{\alpha}_{\text{Mean}}$. This is indeed the case, and is detailed next.

1.6 Small Sample Properties of the Point Estimators

1.6.1 Computation

Having operationalized all three bias–corrected estimators in a fast and accurate fashion, it becomes feasible to conduct a simulation study in order to asses their properties, along with the OLS estimator, and the exact maximum likelihood estimator $\hat{\alpha}_{\text{ML}}$. For even greater time savings, the specific simulation scheme used capitalizes on the fact that the three bias–corrected estimators are one–to–one transformations of the least squares estimator; the details are given in Appendix 1.C. The corresponding Matlab programs are available from the authors.

The median, mean, and mode–adjusted estimators all have to be truncated above at unity because their existence crucially depends on the invertibility of the median, mean, and mode functions. In order for this condition to hold, the respective functions have to be strictly increasing, which was found to be the case when $|\alpha| \leq 1$. To improve comparability, we chose to restrict $\hat{\alpha}_{\text{LS}}$ and $\hat{\alpha}_{\text{ML}}$ as well, which anyway mirrors what would be done in practice when working with economic data for which an explosive process ($\alpha > 1$) is untenable. In doing so, observe that, for α values close to or precisely unity, estimators which adjust by larger amounts will be favored.

1.6.2 Results for a Typical Model

We now discuss in some detail the results for the model with constant and time trend, and $T = 25$ observations (the small sample size being used to help illustrate the differences in the methods; see below for larger T). Figure 1.2 plots the mean bias, median bias and MSE of the various estimators as a function of α, computed at $\alpha = -1, -0.8, \ldots, 0, 0.1, 0.2, \ldots, 0.9, 0.91, 0.92, \ldots, 1$. Regarding mean bias, defined as $\mathrm{E}[\hat{\alpha}] - \alpha$, $\hat{\alpha}_{\mathrm{Mean}}$ is indeed the least biased for all values of α, but still deviates from zero considerably as α approaches unity. Not surprisingly, for virtually the entire parameter space of interest ($\alpha > -0.5$), $\hat{\alpha}_{\mathrm{LS}}$ is the most biased, drastically so as α approaches unity. Regarding median bias, we confirm that $\hat{\alpha}_{\mathrm{Med}}$ is indeed unbiased, while the bias of the other adjusted estimators is not particularly large. (The very small spike in the median bias of $\hat{\alpha}_{\mathrm{Med}}$ near 0.96 is indeed due to use of the SPA for calculation of the relevant cdf. The spike disappears completely for sample sizes $T \geq 35$.)

However, for any estimator to be a useful inferential tool, we require not only that its distribution be centered at the true parameter, but also maximally concentrated around it, as measured naturally by its absolute moments about that value,

$$\mathbb{E}\left[|\hat{\alpha} - \alpha|^d\right], \quad d > 0.$$

As was stressed above, an estimator with less bias does not necessarily perform better with regard to such a measure, because the bias correction procedure itself may increase dispersion. Taking $d = 2$, i.e., the mean squared error, this trade–off is embodied in the well–known decomposition, MSE = BIAS^2 + VARIANCE. In our study, we therefore focus on the MSE as a measure of concentration, as, in combination with the above results regarding bias, it allows us to discern the source of estimation error. In order to verify the robustness of our findings, we also considered the median absolute deviation (MAD) of the estimators, but the results were largely similar and thus are not reported.

For most of the negative α region, the exact MLE performs best, while for $-0.1 \leq \alpha \leq 0.7$, $\hat{\alpha}_{\mathrm{Mode}}$ exhibits the smallest MSE. For $\alpha > 0.7$, $\hat{\alpha}_{\mathrm{Mean}}$ is the best; this result is of particular interest because of the predominance in economic data sets of values of α near unity. It should be noted, however, that $\hat{\alpha}_{\mathrm{Mean}}$ exhibits the highest MSE of all the estimators (including $\hat{\alpha}_{\mathrm{LS}}$) for $-0.6 \leq \alpha \leq 0.4$, while for $0.4 \leq \alpha \leq 0.7$, $\hat{\alpha}_{\mathrm{Mean}}$ has the highest MSE among all the bias–corrected estimators. $\hat{\alpha}_{\mathrm{Med}}$ never achieves the lowest MSE, except at a point near $\alpha = 0.7$, where the MSE of all three bias–corrected estimators cross.

1.6.3 Results for other Parameterizations

It must be kept in mind that the previous discussion pertains only to the specific $\mathbf{X}$ matrix and sample size under consideration. Rather conveniently however, it turns out that the results are qualitatively extremely similar for different sample sizes, $\mathbf{X}$ matrices, and distributional assump-

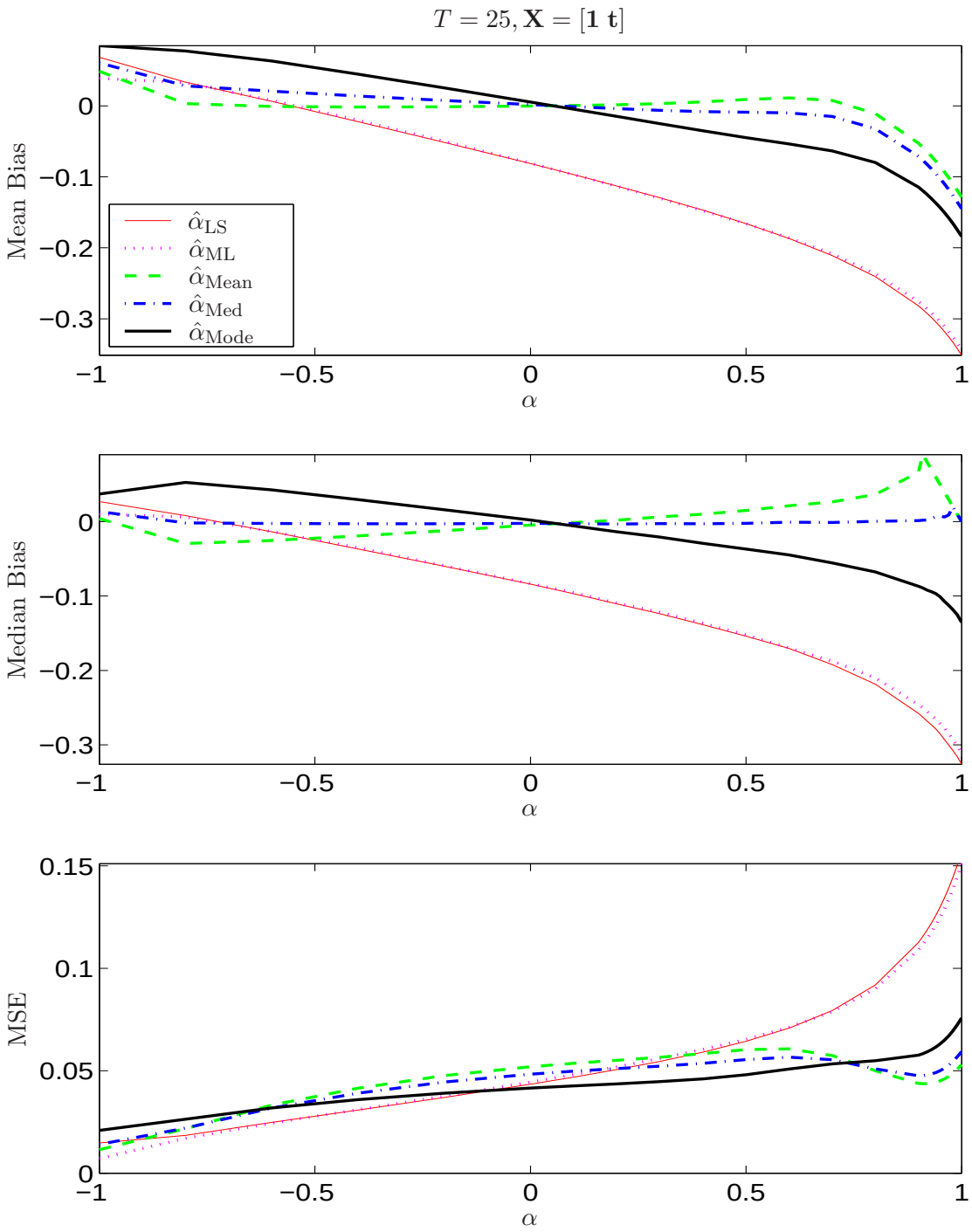

Figure 1.2: Mean bias (top), median bias (middle) and MSE (bottom) of the estimators $\hat{\alpha}_{\text{LS}}$ (thin solid line), $\hat{\alpha}_{\text{ML}}$ (dotted), $\hat{\alpha}_{\text{Mean}}$ (dashed), $\hat{\alpha}_{\text{Med}}$ (dash-dot), and $\hat{\alpha}_{\text{Mode}}$ (thick solid line) based on a model with constant and time trend, with $T = 25$ observations and normal innovations. Bottom graph is truncated; the MSE of $\hat{\alpha}_{\text{LS}}$ increases to 0.16 at $\alpha = 1$.

tions, so that general conclusions can be drawn. This performance is now examined in some more detail.

First consider changing $\mathbf{X}$ to just a column of ones, denoted $\mathbf{X} = \mathbf{1}$. Figure 1.3 shows the results for the MSE (the bias results were very similar to those shown in Figure 1.2 and are omitted). The general shape of the MSE as a function of α is more hump–shaped, but the ranges of α for which a particular estimator is preferred are virtually the same. Noticeable is that the MSE of the bias–corrected estimators using $\mathbf{X} = \mathbf{1}$ does not increase as much as α approaches unity. Also, $\hat{\alpha}_{\text{Mean}}$ now exhibits the highest MSE over an even larger range of α, somewhat more than half the parameter space, and $\hat{\alpha}_{\text{Med}}$ is the second worst for most of the same region.

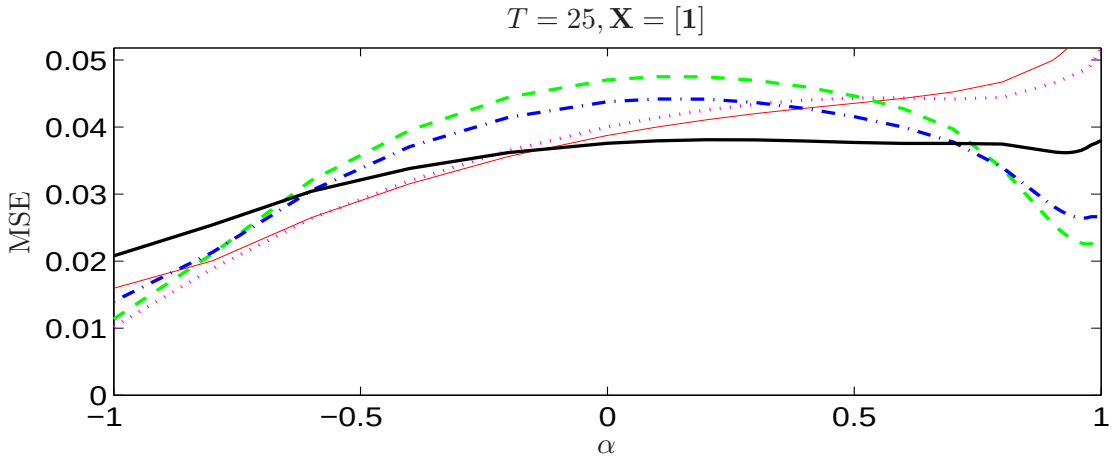

Figure 1.3: Same as the bottom panel in Figure 1.2 but having used only a constant (and no trend) in the regressor matrix.

We now resume use of the constant and trend model with $T = 25$ observations, but consider changing the distributional assumption from normal to Cauchy, which possesses tails much fatter than usually arises in empirical applications in econometrics and serves as a special case of both the Student's t and symmetric stable Paretian distribution.

It should be kept in mind that the bias–corrected estimators are all based on the normal assumption used in the calculation of the distribution of $\hat{\alpha}_{\text{LS}}$ in (1.14). The distribution of quadratic forms in variables other than normal is virtually intractable, though some results on their moments are available; see Roberts (1995), Ullah et al. (1995) and the references therein. In the case of Cauchy innovations, the asymptotic distribution of the OLS estimator involves the ratio of two independent stable Paretian random variables (Davis and Resnick, 1986). However, because we truncate the distributions of the estimators at -1 and 1, the mean bias and MSE are still meaningful statistics.

The results are shown in Figure 1.4. While there are certain differences, the overall behavior of the estimators is still similar to the normal case. For example, $\hat{\alpha}_{\text{Mean}}$ is still approximately unbiased over most of the parameter space, exhibiting an increase in bias as $|\alpha|$ approaches one,

as in the normal case. Estimator $\hat{\alpha}_{\text{Med}}$ is no longer median unbiased, but is approximately so for $\alpha < 0.8$. Interestingly, $\hat{\alpha}_{\text{Mode}}$ is also approximately median–unbiased.

The differences in MSE among the bias–corrected estimators are somewhat less pronounced, although, qualitatively speaking, the envelope of minimum MSE is virtually the same, i.e., $\hat{\alpha}_{\text{LS}}$ is recommended over most of the negative α range, $\hat{\alpha}_{\text{Mode}}$ for $-0.1 \leq \alpha \leq 0.7$ and $\hat{\alpha}_{\text{Mean}}$ for $\alpha > 0.7$.

A similar analysis was conducted using other distributional assumptions including Laplace, Student's t and asymmetric stable Paretian (the latter two with tail indexes such that the mean exists). We chose these distributions (instead of other candidates previously used in similar comparison exercises, such a chi-square or uniform) because their use has become commonplace for capturing the often–observed non–normality of economic and financial data (see, for example, McDonald, 1997; Adler et al., 1998; and Kotz et al., 2001, for a vast array of applications). The bias and MSE results were barely distinguishable from those based on a Gaussian assumption, even when using extremely leptokurtic and asymmetric innovations. It appears that, for small sample sizes, the choice of $\mathbf{X}$ has more of an impact than does—even considerable—deviation from normality, in terms of both fatter tails and/or asymmetry.

To further investigate the robustness of our findings, various other $\mathbf{X}$–matrix specifications were tried, such as (i) addition of boolean (dummy) vectors, as would be used, for example, when working with models with outliers or structural breaks, and (ii) the matrix $\mathbf{E}_k$, specified as the first k eigenvectors of the first order difference matrix, for various values of k. Matrix $\mathbf{E}_k$ is given by $x_{it} = \cos[(2t-1)\pi(i-1)/(2T)]$ (see Durbin and Watson, 1971), and is a useful benchmark because these vectors tend to mimic the behavior of economic time series with seasonal and cyclical–type behavior (Dubbelman et al., 1978; King, 1985, p. 32). For all $\mathbf{X}$–matrices considered, the ranges for which the respective estimators perform best were virtually identical.

Finally, to see the effect of sample size, Figure 1.5 shows the MSE results for the constant–trend model with normal innovations, but now using $T = 75$ observations. As expected, the MSE decreases for all estimators. The α–ranges and estimators corresponding to the minimum MSE envelope are again virtually the same, but now the difference in MSE of the bias–corrected estimators is far less pronounced. This is expected, because, for $|\alpha| < 1$, $\sqrt{T}(\hat{\alpha}_{\text{LS}} - \alpha) \overset{\text{asy}}{\sim} \text{N}(0, 1-\alpha^2)$, for which the mean, median and mode coincide. The shape of the MSE curve in Figure 1.5 is also much closer to $(1-\alpha^2)/T$ than those corresponding to $T = 25$. Also, because the asymptotic distribution is less accurate for a given sample size as α approaches one, the discrepancy in MSE shown in Figure 1.5 increases as α approaches one (with $\hat{\alpha}_{\text{Mean}}$ exhibiting the lowest MSE).

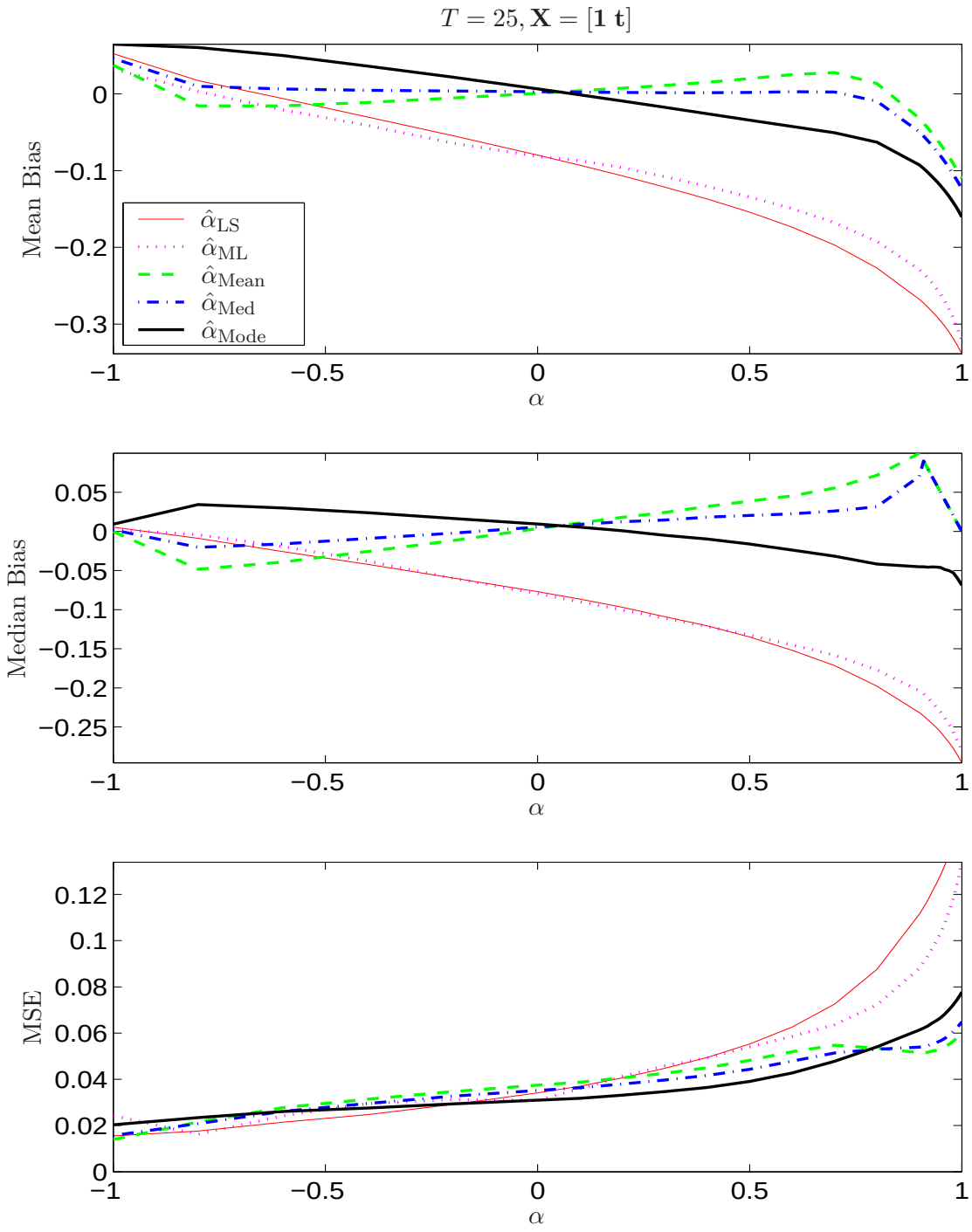

Figure 1.4: Same as Figure 1.2 but having used Cauchy innovations

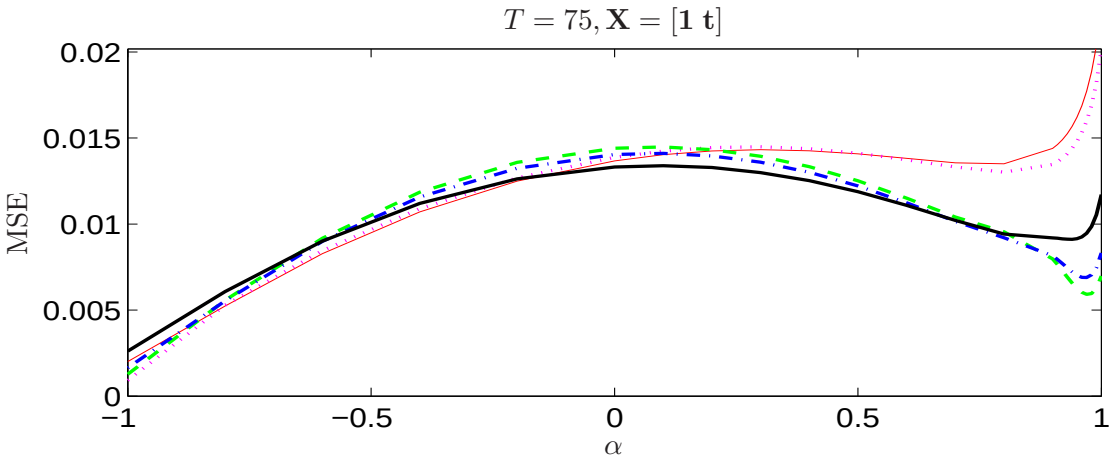

Figure 1.5: Same as the bottom panel of Figure 1.2 but using $T = 75$

1.7 Conclusions

The median–unbiased point estimator suggested by Andrews (1993) for the first order autoregressive coefficient is statistically well–motivated and also computationally feasible, owing to the tabulated values provided by Andrews. Nevertheless, its implementation based on these tables was restricted to models with either no intercept, intercept, or intercept and trend, while many models will require different exogenous regressors, including dummies to pick up structural breaks or outliers, thus requiring time–consuming custom calculations. This paper uses a saddlepoint approximation to the required distribution function, rendering such custom calculations feasible as a routine task. The mean–adjusted estimator of Tanizaki (2000) suffers from a similar computational burden, owing to the use of simulation in the construction of the mean function. We have shown how this time–consuming process can be replaced with a fast, exact calculation.

The availability and high accuracy of the saddlepoint approximation facilitates other computationally intensive estimation methods. In particular, we propose a new estimator which uses the mode as a measure of central tendency. This estimator takes under a second to compute (when using the density saddlepoint approximation), and does not appear to have been entertained in any statistical modelling context previous to this study.

With all the estimators computable in a fast and reliable fashion, a simulation study comparing their small–sample properties becomes feasible. With respect to MSE, we demonstrate that $\hat{\alpha}_{\text{Mode}}$ is superior for the large part of the parameter space, $-0.1 < \alpha < 0.7$, while $\hat{\alpha}_{\text{Mean}}$ is shown to exhibit lower MSE for the important region $0.7 < \alpha < 1$.

Perhaps the most important observation regarding these results on MSE is that they appear to hold almost exactly irrespective of the choice of exogenous regressors, sample size, and innovation distribution assumption. Concretely, this means that an estimator with relatively lowest MSE (of the estimators entertained herein) can be selected, assuming correct prior opinion on whether

$\alpha < -0.1$, $-0.1 < \alpha < 0.7$ or $0.7 < \alpha < 1$, the latter being a common choice.

Software for Matlab is available from the authors (i) to compute the estimators discussed herein for a given data set, and (ii) to determine the properties of the point estimators over a grid of values of the autoregressive parameter for any exogenous regressor matrix and choice of innovation assumption (included are normal, Student's t, stable Paretian and Laplace, though others are easily incorporated). This can be used to help decide on a suitable point estimator to report based on the observed data, although, as discussed above, our simulations show that the optimal choice of estimator is virtually invariant to the regressor matrix, sample size, and innovations assumption.

Given the ease of computation, one could entertain a two step procedure when faced with no prior information: First estimate α with, say, $\hat{\alpha}_{\text{Mode}}$, which exhibits good overall MSE performance, and, based on it, the estimator with lowest MSE is selected to deliver the final estimate. The small sample performance of such a procedure might be worth investigating, although, realistically, most researchers will have, to some extent, a prior on α.

The estimation techniques developed herein could be applied to models for panel data; for example, Phillips and Sul (2003) make use of a median unbiased estimator in the context of dynamic panel models. We leave this extension for future analysis.

Appendix 1.A Proof of the Independence of $\hat{\alpha}_{\text{LS}}$ and $\boldsymbol{\beta}$

We need to show that $\hat{\alpha}_{\text{LS}}$ in (1.5) is independent of $\boldsymbol{\beta}$ for any exogenous regressor matrix $\mathbf{X}$, as remarked without proof by Andrews (1993, p. 146). Such a proof could proceed in (at least) two ways. One is based on the observation that the OLS estimator is a function of Y only through a maximal invariant which has distribution free of β, see, e.g., Dufour and King (1991). A second way is to use a singular value decomposition of $\mathbf{X}$, which also allows linearly dependent regressors, and is now demonstrated.

First note that the computation of (1.5) requires the $T \times 2k$ matrix $\mathbf{Z}$ to be of full column rank which is not always satisfied, e.g., if the regressors include a constant, a constant and a linear trend, or a specific combination of impulse and step dummies. Typically, the linearly dependent columns are "calculated out" by hand, such as in Andrews (1993) for the constant and trend model.

To make the proof valid for all $r = \text{rank}(\mathbf{Z}) \leq 2k$, we apply a singular value decomposition (SVD) to $\mathbf{Z}$,

$$\mathbf{Z} = \mathbf{Q}\mathbf{W}\mathbf{V}',$$

where $\mathbf{Q}$ and $\mathbf{V}$ are $T \times r$ and $2k \times r$ matrices, respectively, of full column rank r and $\mathbf{W}$ is an $r \times r$ diagonal matrix of full rank. Moreover, $\mathbf{Q}'\mathbf{Q} = \mathbf{V}'\mathbf{V} = \mathbf{I}_r$. Clearly, only r different

parameters in $\boldsymbol{\gamma}$ are identified. Defining $\tilde{\mathbf{Z}} = \mathbf{QW}$ and $\tilde{\boldsymbol{\gamma}} = \mathbf{V}'\boldsymbol{\gamma}$, we can rewrite (1.3) as

$$\mathbf{Y}_T = \mathbf{Y}_{T-1}\alpha + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\gamma}} + \mathbf{U}_T.$$

The linearly dependent columns of $\mathbf{Z}$ are effectively removed by means of the SVD of $\mathbf{Z}$. The OLS estimator for α can now be obtained as in (1.5) with $\mathbf{M}$ replaced by $\mathbf{M} = \mathbf{I}_T - \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}' = \mathbf{I}_T - \mathbf{QQ}'$.

Following Andrews (1993), to show that $\hat{\alpha}_{\mathrm{LS}}$ does not depend on the value of $\boldsymbol{\beta}$, it suffices to show that the first–step residuals $\mathbf{MY}_T$ and $\mathbf{MY}_{T-1}$ do not depend on $\boldsymbol{\beta}$. Because of $\mathbf{MY}_T = \mathbf{MX}_T\boldsymbol{\beta} + \mathbf{MY}_T^{\ell}$ and $\mathbf{MY}_{T-1} = \mathbf{MX}_{T-1}\boldsymbol{\beta} + \mathbf{MY}_{T-1}^{\ell}$, where $\mathbf{Y}_T^{\ell}$ and $\mathbf{Y}_{T-1}^{\ell}$ are defined analogously to $\mathbf{Y}_T$ and $\mathbf{Y}_{T-1}$, this amounts to showing that $\mathbf{MX}_T = \mathbf{0}$ and $\mathbf{MX}_{T-1} = \mathbf{0}$ because neither $\mathbf{MY}_T^{\ell}$ nor $\mathbf{MY}_{T-1}^{\ell}$ depend on $\boldsymbol{\beta}$. Partitioning $\mathbf{V}'$ into the first k and the last k columns $\mathbf{V}_1'$ and $\mathbf{V}_2'$, respectively, we obtain $\mathbf{Z} = [\mathbf{X}_T, \mathbf{X}_{T-1}] = [\mathbf{QWV}_1', \mathbf{QWV}_2']$ and, thus, $\mathbf{X}_T = \mathbf{QWV}_1'$ and $\mathbf{X}_{T-1} = \mathbf{QWV}_2'$. Now, $\mathbf{MX}_T = (\mathbf{I}_T - \mathbf{QQ}')\mathbf{QWV}_1' = (\mathbf{Q} - \mathbf{Q})\mathbf{WV}_1' = \mathbf{0}$ and $\mathbf{MX}_T = (\mathbf{I}_T - \mathbf{QQ}')\mathbf{QWV}_2' = (\mathbf{Q} - \mathbf{Q})\mathbf{WV}_2' = \mathbf{0}$, which proves the invariance of $\hat{\alpha}_{\mathrm{LS}}$ with respect to $\boldsymbol{\beta}$. The invariance of $\hat{\alpha}_{\mathrm{LS}}$ with respect to σ^2 (and to Y_0 if $\alpha = 1$) can be proved as outlined by Andrews (1993). Consequently, we can assume $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$ in the following.

Defining the selection matrices $\mathbf{D}_T = [\mathbf{0} \mid \mathbf{I}_T]$ and $\mathbf{D}_{T-1} = [\mathbf{I}_T \mid \mathbf{0}]$, we now have $\mathbf{MY}_T = \mathbf{MD}_T\mathbf{Y}^{\ell}$ and $\mathbf{MY}_{T-1} = \mathbf{MD}_{T-1}\mathbf{Y}^{\ell}$, where $\mathbf{Y}^{\ell} = [Y_0^{\ell}, (\mathbf{Y}_T^{\ell})']'$. Substituting this into (1.5) yields

$$\hat{\alpha}_{\mathrm{LS}} = \frac{(\mathbf{Y}^{\ell})'\mathbf{D}_{T-1}'\mathbf{MD}_T'\mathbf{Y}^{\ell}}{(\mathbf{Y}^{\ell})'\mathbf{D}_{T-1}'\mathbf{MD}_{T-1}'\mathbf{Y}^{\ell}} = \frac{\mathbf{U}'\mathbf{R}'\mathbf{D}_{T-1}'\mathbf{MD}_T\mathbf{RU}}{\mathbf{U}'\mathbf{R}'\mathbf{D}_{T-1}'\mathbf{MD}_{T-1}\mathbf{RU}} = \frac{\mathbf{U}'\mathbf{AU}}{\mathbf{U}'\mathbf{BU}}, \tag{1.14}$$

where $\mathbf{A}$ and $\mathbf{B}$ are so defined and $\mathbf{Y}^{\ell} = \mathbf{RU}$ for $\mathbf{U} = [U_0, \ldots, U_T] \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{T+1})$, with

$$\mathbf{R} = \mathbf{R}(\alpha) = \begin{bmatrix} b & 0 & 0 & \cdots & 0 & 0 \\ b\alpha & 1 & 0 & \cdots & 0 & 0 \\ b\alpha^2 & \alpha & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b\alpha^T & \alpha^{T-1} & \alpha^{T-2} & \cdots & \alpha & 1 \end{bmatrix},$$

$b = \left(1 - \alpha^2\right)^{-1/2}$ if $\alpha \in (-1, 1)$ and zero if $\alpha = 1$. If there are no exogenous regressors, set $\mathbf{M} = \mathbf{I}$ and all conditioning on $\mathbf{X}$ is replaced by conditioning on T.

Appendix 1.B Saddlepoint Approximation to the cdf and pdf of $\hat{\alpha}_{\mathrm{LS}}$

For given α, the cdf of $\hat{\alpha}$ can be expressed as

$$\Pr(\hat{\alpha}_{\mathrm{LS}} \leq c) = \Pr\left(\mathbf{U}'\left(\mathbf{A}/2 + \mathbf{A}'/2 - c\mathbf{B}\right)\mathbf{U} \leq 0\right) = \Pr\left(\mathbf{U}'\mathbf{WU} \leq 0\right), \tag{1.15}$$

with symmetric matrix $\mathbf{W} = \mathbf{W}(\alpha, c)$ so defined. From the principle axis theorem,

$$\Pr\left(\mathbf{U}'\mathbf{W}\mathbf{U} \leq 0\right) = \Pr\left(\sum_{i=1}^{w} \lambda_i \chi_i^2(1,0) \leq 0\right) = \Pr\left(S \leq 0\right), \tag{1.16}$$

where S is the so–defined weighted sum, $w = \text{rank}(\mathbf{W})$, each $\chi_i^2(1,0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the λ_i are the eigenvalues of $\mathbf{W}$. Let $K = K_S$ be the cumulant generating function of S, given by $K_S(s) = \frac{1}{2}\sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.

The general result of Lugannani and Rice (1980) can be directly applied to S to yield the saddlepoint approximation to its cdf; it is

$$\tilde{F}_S(x) = \Phi(\hat{w}) + \phi(\hat{w})\left\{\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right\}, \qquad x \neq \mathrm{E}[S], \tag{1.17}$$

where $\hat{w} = \text{sgn}(\hat{s})\sqrt{2\hat{s}x - 2K(\hat{s})}$, $\hat{u} = \hat{s}\sqrt{K'(\hat{s})}$, Φ and ϕ are the cdf and pdf of the standard normal distribution, respectively, and $\hat{s}$ is the (unique) saddlepoint which satisfies $x = K_S'(\hat{s})$. This needs to be numerically solved. Daniels (1987) derived the next term in the expansion (1.17), given by

$$\hat{F}_S(x) = \tilde{F}_S(x) - \phi(\hat{w})\left\{\hat{u}^{-1}\left(\frac{\hat{\kappa}_4}{8} - \frac{5}{24}\hat{\kappa}_3^2\right) - \hat{u}^{-3} - \frac{\hat{\kappa}_3}{2\hat{u}^2} + \hat{w}^{-3}\right\}, \qquad x \neq \mathrm{E}[S], \tag{1.18}$$

where $\hat{\kappa}_i = K^{(i)}(\hat{s})/[K''(\hat{s})]^{i/2}$. Use of (1.18) was found to be virtually always more accurate than (1.17) and is the method used in all calculations above.

While limiting expressions for the case $x = \mathrm{E}[S]$ do exist, numerical problems will arise for values of x sufficiently close to $\mathrm{E}[S]$. This is easily circumvented in practice by linear interpolation of values in a small neighborhood of x.

A second order SPA to the pdf of $\hat{\alpha}_{\text{LS}}$ is developed in Lieberman (1994b). With $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}(\alpha, c)$ and $v_i, i = 1, \ldots, w$, as previously defined, the first-order approximation is given by

$$\tilde{f}_{\hat{\alpha}_{\text{LS}}}(c) = \frac{\text{tr}[(\mathbf{I} - 2\hat{s}\mathbf{W})^{-1}\mathbf{B}]\exp\left\{\frac{1}{2}\sum_{i=1}^{w}\ln v_i\right\}}{\sqrt{4\pi\sum_{i=1}^{w}(\lambda_i v_i)^{-2}}}, \tag{1.19}$$

where $\hat{s}$ denotes the same unique saddlepoint as is used in approximating the cdf.

Expression (1.19) is the leading term in an asymptotic expansion; the second order approximation is given by

$$\hat{f}_{\hat{\alpha}_{\text{LS}}}(c) = \tilde{f}_{\hat{\alpha}_{\text{LS}}}(c)\left(1 - \frac{2\,\text{tr}\left(\mathbf{K}^2\mathbf{L}\right)}{\text{tr}\left(\mathbf{L}\right)\text{tr}\left(\mathbf{K}^2\right)} + \frac{3\,\text{tr}\left(\mathbf{K}^4\right)}{2\left(\text{tr}\,\mathbf{K}^2\right)^2} + \frac{2\,\text{tr}\left(\mathbf{K}\mathbf{L}\right)\text{tr}\left(\mathbf{K}^3\right)}{\text{tr}\left(\mathbf{L}\right)\left(\text{tr}\,\mathbf{K}^2\right)^2} - \frac{5\left(\text{tr}\,\mathbf{K}^3\right)^2}{3\left(\text{tr}\,\mathbf{K}^2\right)^3}\right), \tag{1.20}$$

where $\mathbf{K} = \hat{\mathbf{H}}^{-1}\mathbf{W}$, $\mathbf{L} = \hat{\mathbf{H}}^{-1}\mathbf{B}$ and $\hat{\mathbf{H}}^{-1} = \hat{\mathbf{H}}^{-1}(\hat{s}) = \mathbf{I} - 2\hat{s}\mathbf{W}$. Calculations reveal that the expression in Lieberman's Equation (5) contains a misprint; it is correct as given here.

Appendix 1.C Simulation Methodology

All three bias–corrected estimators are one–to–one transformations of the least squares estimator, say $\hat{\alpha}_{\text{BC}} = m_{\text{BC}}^{-1}(\hat{\alpha}_{\text{LS}})$, where BC denotes the respective method of bias correction, i.e., BC $\in \{\text{Mean}, \text{Med}, \text{Mode}\}$, and $m_{\text{BC}}^{-1}(\hat{\alpha}_{\text{LS}})$ is the inverse mean, median, and mode function, respectively. Of course, $m_{\text{BC}}^{-1}(\hat{\alpha}_{\text{LS}})$ is not available analytically, but can be computed by numerical methods. Thus, for a given sample size and $\mathbf{X}$–matrix, the distribution of the adjusted estimators could be calculated, from which properties of interest such as the median and moments (for the bias and MSE) could be obtained by integration. Alternatively, as each of the three estimators takes under a second to compute on a modern PC, a direct, brute–force simulation exercise is also feasible. We use a combination of these two methods, which involves simulation, but capitalizes on the one–to–one property of the estimators. The algorithm is such that only $\hat{\alpha}_{\text{LS}}$ needs to be replicated via simulation, resulting in a ten–fold decrease in computation time compared to direct simulation.

The first step is to compute $m_{\text{BC}}^{-1}(\hat{\alpha}_{\text{LS}})$ for a grid of values. The choice of grid is decisive if both high accuracy and high speed (i.e., minimal computational effort) are desired. The naive approach of using an equally spaced grid with a pre-specified number of points turns out to be highly inefficient, as $m_{\text{BC}}^{-1}(\hat{\alpha}_{\text{LS}})$ is not linear over the whole range, but rather begins to curve at a point near unity (the exact location depending on $\mathbf{X}$, sample size and choice of estimator). The solid line in Figure 1.6 illustrates m^{-1} for $\hat{\alpha}_{\text{Mean}}$, based on the intercept and time trend model with $T = 25$. Because of the nonlinearity, it is advantageous to construct a dynamic grid selection procedure. Pseudocode for the recursive scheme used for all calculations presented in this paper is presented below.

FUNCTION $[\mathbf{x}, \mathbf{m}] = \text{makegrid}\big(\underline{x}, \overline{x}, m_{\text{BC}}^{-1}(\underline{x}), m_{\text{BC}}^{-1}(\overline{x})\big)$
Set $x = \frac{1}{2}(\underline{x} + \overline{x})$;
Compute $\hat{m}_{\text{BC}}^{-1}(x) := m_{\text{BC}}^{-1}(\underline{x}) + \frac{1}{2}\left[m_{\text{BC}}^{-1}(\overline{x}) - m_{\text{BC}}^{-1}(\underline{x})\right]$;
IF $\Big(\big(m_{\text{BC}}^{-1}(x) - \hat{m}_{\text{BC}}^{-1}(x) > \varepsilon \text{ AND } \overline{x} - x > S_{\text{Min}}\big) \text{ OR } \overline{x} - x > S_{\text{Max}}\Big)$
  $[\underline{\mathbf{x}}, \underline{\mathbf{m}}] = \text{makegrid}(\underline{x}, x, m_{\text{BC}}^{-1}(\underline{x}), m_{\text{BC}}^{-1}(x))$;
  $[\overline{\mathbf{x}}, \overline{\mathbf{m}}] = \text{makegrid}(x, \overline{x}, m_{\text{BC}}^{-1}(x), m_{\text{BC}}^{-1}(\overline{x}))$;
ELSE
  $\underline{\mathbf{x}} = \{\}; \overline{\mathbf{x}} = \{\}; \underline{\mathbf{m}} = \{\}; \overline{\mathbf{m}} = \{\}$;
END
$\mathbf{x} = [\underline{\mathbf{x}}, x, \overline{\mathbf{x}}]$; $\mathbf{m} = [\underline{\mathbf{m}}, m_{\text{BC}}^{-1}(x), \overline{\mathbf{m}}]$;

Calling the above function with $\big(\underline{x}, \overline{x}, m_{\text{BC}}^{-1}(\underline{x}), m_{\text{BC}}^{-1}(\overline{x})\big) = \big(-1, 1, m_{\text{BC}}^{-1}(-1), m_{\text{BC}}^{-1}(1)\big)$ returns a list of grid points and the corresponding function values. Parameter ε dictates the spacing of the grid, while S_{Min} and S_{Max} define lower and upper bounds on the spacing. The values of $\varepsilon = 10^{-4}$, $S_{\text{Min}} = 10^{-4}$ and $S_{\text{Max}} = 10^{-1}$ were found to be acceptable for all cases considered in

the paper and are the default program values. Using these for the grid construction, and 10,000 replications in the subsequent simulation, a run with $T = 25$ takes about 18 minutes on a typical PC with a 2.5 GHz Pentium 4 processor. Run times are also practically linear in T for $T < 100$; for $T = 50$, 38 minutes were required. For $\hat{\alpha}_{\text{Mean}}$ with $T = 25$, the resulting set of about 200 grid points is shown as slashes in Figure 1.6.

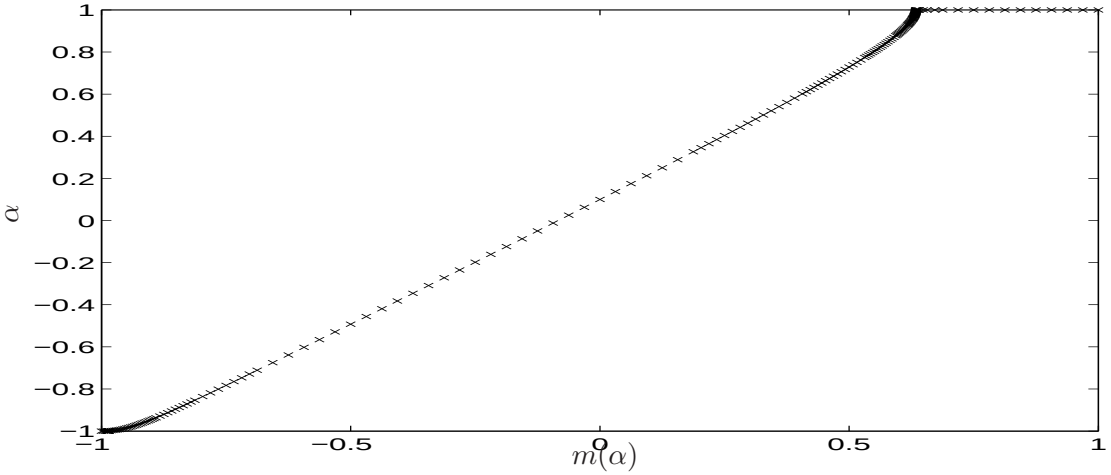

Figure 1.6: Inverse mean function for a model with constant and time trend, $T = 25$.

The second step involves the following: (i) simulate a time series from (1.1), (ii) calculate $\hat{\alpha}_{\text{LS}}$, and (iii) based on $\hat{\alpha}_{\text{LS}}$, use linear interpolation from the grid obtained in the first step to obtain the corresponding bias-corrected estimators. As (i), (ii) and (iii) are all numerically trivial, this second step is extremely fast, enabling use of a very large number of replications (we used 10,000), so that the inherent variation arising from simulation can be effectively eliminated. Thus, for any given $\mathbf{X}$ matrix, the mean bias, median bias and the MSE of the three bias–adjusted estimators can be routinely computed over a grid of α–values.

Manuscript 2

Approximately Exact Inference in Dynamic Panel Models: a QUEST for Unbiasedness

2.1 Introduction

Dynamic panel data (DPD) models receive a considerable amount of attention in both the theoretical and applied literature (see, for example, the references in Arellano, 2003). Due to its tractability and wide applicability, the first–order DPD model is by far the most popular. Since the seminal work of Anderson and Hsiao (1981), the literature has mainly focused on generalized method of moments (GMM) procedures for its estimation. The first–difference GMM estimator introduced by Arellano and Bond (1991) is now one of the standard procedures used in empirical applications. Ahn and Schmidt (1995) exploit additional moment conditions in the presence of exogenous variables. More recently, Kruiniger (2002) examines various estimators, under different specifications of the individual effects, and derives conditions of consistency. A recent addition to this strand of literature is Moon and Phillips (2004).

One of the reasons why GMM procedures enjoy such popularity is the *"incidental parameters"* problem and associated asymptotic bias, which was first discussed by Neyman and Scott (1948) and is pertinent to the least squares– and maximum likelihood estimators in (fixed effects) DPD models with a large number of individuals N (Nickell, 1981). GMM estimators, on the other hand, rely on large-N asymptotics, but may be severely biased in samples with a small number of individuals, as are common in macroeconomic panels. This problem of bias has led to the development of alternative estimators; examples include the minimum distance estimator of Hsiao et al. (2002), and the recursive mean adjustment technique of Choi et al. (2004).

In some recent publications, attempts have been made at correcting the bias of the least squares estimator. The relevant literature can roughly be divided into two strands, based on the method used to obtain the requisite bias function. One group of authors (Phillips and Sul, 2003; Gouriéroux et al., 2006; Everaert and Pozzi, 2007) takes a simulation approach, which, while being valid in small samples and allowing for quite general model specifications, is computationally demanding and susceptible to numerical inaccuracies. The other class of papers (Kiviet, 1995; Hahn and Kuersteiner, 2002; Bun and Carree, 2005) relies on asymptotic results, typically as the number of individuals tends to infinity. While not suffering from the aforementioned drawbacks, these methods require tedious custom calculations for each model to be analyzed, and, owing to their asymptotic nature, tend to break down in small samples. As Hahn and Kuersteiner (2002, p. 1647) note,

> Unfortunately, our bias-corrected estimator does not completely remove the bias. This suggests that an even more careful small sample analysis based on higher order expansions of the distribution might be needed to account for the entire bias.

The present manuscript is geared at providing such an analysis: it develops the concept of quantile unbiased estimation, or QUEST, as a general method for median unbiased point estimation and the construction of exact confidence intervals, and applies it to the first–order DPD model. Unlike previous work in the field, our approach is built around the exact maximum

likelihood — rather than least squares — estimator of the autoregressive coefficient, which allows us to construct a saddlepoint approximation to the tail probabilities of the requisite estimating functions. Use of our otherwise exact inferential procedure in conjunction with the proposed saddlepoint approximation gives rise to the seemingly contradictory nomenclature *approximately exact* inference, a term originally coined by Strawderman and Wells (1998). The method obviates the need for burdensome simulations, yet simultaneously allows us to maintain small-sample validity and the flexibility of arbitrary sets of exogenous regressors. Another distinguishing feature of our approach is that it explicitly allows for non-homogenous individual error variances. This is in stark contrast with most other works on exact inference in DPD models, including those that make use of saddlepoint approximations (Paige and Trindade, 2006; Perera et al., 2006), the likely explanation being that the relevant distribution function has previously been intractable.

The remainder of this paper is organized as follows: Section 2.2 presents the general method for point and interval estimation. Section 2.3 introduces the dynamic panel model. Sections 2.4 and 2.5 apply the estimation methodology to the least squares and maximum likelihood estimation of the model, respectively. Section 2.6 contains numerical results. Section 2.7 concludes.

2.2 Quantile Unbiased Estimation

This section develops a general procedure for conducting exact inference in models that allow the estimator of the parameter of interest to be defined as the unique root of an estimating function. The method generalizes the approach of Andrews (1993) and is related to the adjusted profile likelihood of McCullagh and Tibshirani (1990). In contrast to the latter, our approach uses quantiles, rather than moments, of the distribution. This has two advantages: i) under certain conditions, the resulting estimator is *exactly* median unbiased, as opposed to *approximately* mean unbiased, ii) it facilitates construction of confidence intervals.

Consider a parametric model $\{\boldsymbol{\Upsilon}, \boldsymbol{\theta}\}$, where $\boldsymbol{\Upsilon}$ is the data, $\boldsymbol{\theta}' = (\theta, \boldsymbol{\delta}')$, $\theta \in [\underline{\theta}, \overline{\theta}]$ is the scalar parameter of interest, $\boldsymbol{\delta}$ is a (possibly empty) set of nuisance parameters, and $\underline{\theta}$, $\overline{\theta}$ are possibly infinite. Consider an estimator of θ defined as the (unique) root of a continuously differentiable estimating function $\eta(\theta, \boldsymbol{\Upsilon})$ that does not involve $\boldsymbol{\delta}$, i.e., $\hat{\theta}(\boldsymbol{\Upsilon})$ solves the estimating equation

$$\hat{\theta}(\boldsymbol{\Upsilon}) = \begin{cases} \underline{\theta}, \text{ if } \eta(\underline{\theta}, \boldsymbol{\Upsilon}) < 0, \\ \overline{\theta}, \text{ if } \eta(\overline{\theta}, \boldsymbol{\Upsilon}) > 0, \\ c : \eta(c, \boldsymbol{\Upsilon}) = 0, \quad \text{otherwise.} \end{cases} \tag{2.1}$$

Let $\Pr_{\boldsymbol{\theta}}(B)$ and $\text{Med}_{\boldsymbol{\theta}}(X)$ denote the probability of B and the median of X if the true parameter is $\boldsymbol{\theta}$, respectively. In analogy to the notion of (mean) unbiased estimating functions, it is natural to call an estimating function η *median unbiased* if

$$\text{Med}_{\boldsymbol{\theta}}\left(\eta(\theta, \boldsymbol{\Upsilon})\right) = 0.$$

More generally, if for a fixed value $q \in (0,1)$, η satisfies

$$\Pr_{\boldsymbol{\theta}}\left(\eta(\theta, \boldsymbol{\Upsilon}) \leq 0\right) = q, \tag{2.2}$$

we shall refer to it (and to the corresponding estimating equation (2.1)) as $100q\%$ *quantile unbiased*.

Our motivation to consider quantile unbiased estimation is the following. It is well known that mean unbiased estimating functions do not, in general, lead to mean unbiased estimators; if, however, we denote by $\mathcal{C}(\boldsymbol{\Upsilon}) \equiv \{c \in (\underline{\theta}, \bar{\theta}) : \eta(c, \boldsymbol{\Upsilon}) = 0\}$ and make the additional assumption

$$\begin{aligned} \left|\mathcal{C}(\boldsymbol{\Upsilon})\right| &\leq 1 \quad \text{almost surely, and} \\ \forall c \in \mathcal{C}(\boldsymbol{\Upsilon}), \frac{\mathrm{d}}{\mathrm{d}c}\eta(c, \boldsymbol{\Upsilon}) &< 0 \quad \text{almost surely,} \end{aligned} \tag{A1}$$

then if η is $100q\%$ quantile unbiased, it follows that

$$q = \Pr_{\boldsymbol{\theta}}\Big(\eta(\theta, \boldsymbol{\Upsilon}) \leq 0\Big) = \Pr_{\boldsymbol{\theta}}\Big(\hat{\theta}(\boldsymbol{\Upsilon}) \leq \theta\Big), \tag{2.3}$$

i.e., $\hat{\theta}$ is a $100q\%$ quantile unbiased estimator for θ. In particular, if η satisfies (2.2) with $q = 0.5$, then its unique root is a median unbiased estimator of θ, while with $q = (1 \pm \tau)/2$, it constitutes the left (right) endpoint of an equal–tails $100\tau\%$ confidence interval.

The following proposition shows how a quantile unbiased estimating function can be constructed for any value of $q \in (0,1)$, and will serve as the main tool in the construction of our estimators. Here and in the sequel, the dependence of η on the data will not generally be made explicit; rather, if $\boldsymbol{\Upsilon}$ appears explicitly, then $\eta(c, \boldsymbol{\Upsilon})$ will be understood as the (observed) sample value of the corresponding statistic.

Proposition 1. *Let $\eta(c) : (\underline{\theta}, \overline{\theta}) \mapsto \mathbb{R}$ be a continuously differentiable estimating function for θ. Assume that, for all c, its distribution function is constant in $\boldsymbol{\delta}$ and denote it by $F_{\eta(c)}(\cdot; \theta)$. Then*

$$\eta^*(c) \equiv \eta(c) - F^{-1}_{\eta(c)}(q; c) \tag{2.4}$$

is a $100q\%$ quantile unbiased estimating function for θ. If, in addition, η^* satisfies (A1), then its unique root is a $100q\%$ quantile unbiased estimator for θ.

Proof. For all values of c,

$$\Pr_c\Big(\eta^*(c) \leq 0\Big) = \Pr_c\Big(\eta(c) \leq F^{-1}_{\eta(c)}(q; c)\Big) = F_{\eta(c)}\Big(F^{-1}_{\eta(c)}(q; c); c\Big) = q,$$

which, in particular, also holds for $c = \theta$, i.e., η^* satisfies (2.2). The second assertion follows at once from (2.3). ■

The root of the unbiased estimating function (2.4), say $\hat{\theta}_q$, can also be expressed as

$$c : \Pr_c\left(\eta(c) \leq \eta(c, \boldsymbol{\Upsilon})\right) = q, \tag{2.5}$$

which will be convenient for our purposes as it obviates the need to calculate the inverse distribution function appearing in (2.4). It is important to note that in (2.5), c occurs both as the argument of the estimating function and as the (hypothesized) true parameter.

We close this section with a few remarks concerning related schemes of bias correction. Firstly, if estimator $\hat{\theta}$ can be expressed in closed form, then it can be written as the solution to

$$\hat{\theta}(\boldsymbol{\Upsilon}) - c = 0,$$

and $\hat{\theta}_q$ solves

$$c : \Pr_c\left(\hat{\theta} \leq \hat{\theta}(\boldsymbol{\Upsilon})\right) = q.$$

In this special case, our technique yields the same estimator as that used by, e.g., Andrews (1993) and Phillips and Sul (2003). Their requirement that the quantile function of $\hat{\theta}$ be strictly increasing in θ translates into our assumption (A1). As noted by Andrews (1993), it is not apparent how this can be formally proven. However, for our model, numerical results appear to confirm this assumption.

Secondly, it appears natural to construct another bias-corrected point estimator by replacing $F^{-1}_{\eta(c)}(q; c)$ in equation (2.4) by $\mathbb{E}_c\left[\eta(c)\right]$, i.e., the expected value of $\eta(c)$ if the true parameter is c, thus giving rise to the mean unbiased estimating function

$$\eta^{**}(c) \equiv \eta(c) - \mathbb{E}_c\left[\eta(c)\right].$$

This is the idea behind the adjusted profile likelihood of McCullagh and Tibshirani (1990), except that here, we are concerned with a general estimation function that need not necessarily be a profile score function. We shall refer to the resulting estimator as *mean adjusted* and denote it by $\hat{\theta}_{\text{Mean}}$. If the estimator in question is closed-form, $\hat{\theta}_{\text{Mean}}$ solves

$$c : \hat{\theta}(\boldsymbol{\Upsilon}) - c - \mathbb{E}_c\left[\hat{\theta} - c\right] = 0, \tag{2.6}$$

which is the nonlinear–bias–correcting estimator of MacKinnon and Smith (1998). A similar idea underlies the indirect inference procedure of Gourieroux et al. (1993).

2.3 The Model

We consider a first–order DPD model, with or without fixed effects. For each of the $N \in \mathbb{N}_+$ individuals, the model is characterized by an observed and a latent equation, given respectively

by

$$\begin{aligned} y_{i,t} &= \mathbf{x}_{i,t}'\boldsymbol{\beta} + y_{i,t}^{\ell}, \quad t \in \{0, \ldots, T\}, \\ y_{i,t}^{\ell} &= \alpha y_{i,t-1}^{\ell} + u_{i,t}, \quad t \in \{1, \ldots, T\}, \end{aligned} \tag{2.7}$$

where $\alpha \in (-1, 1]$, $\mathbf{x}_{i,t} = (x_{i,t}^1, \ldots, x_{i,t}^k)'$ is a vector of regressors with $k < NT$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, the error components $u_{i,t} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_i^2)$, and each initialization $y_{i,0}^{\ell} \sim \mathrm{N}\left(0, \frac{\sigma_i^2}{1-\alpha^2}\right)$ if $\alpha \in (-1, 1)$ and an arbitrary constant or random variable if $\alpha = 1$. In order to keep the notation consistent throughout the paper, we define $\boldsymbol{\beta}$ in equation (2.7) to contain the coefficients for *all* individuals, and we set $x_{i,t}^k = 0 \; \forall i, t$ if regressor x^k does not appear in equation i.

In matrix form, the observed and latent panels become

$$\begin{aligned} \mathbf{Y}_0 &= \mathbf{X}_0\boldsymbol{\beta} + \mathbf{Y}_0^{\ell} \quad \text{and} \\ \mathbf{Y}^{\ell} &= \alpha\mathbf{Y}_{-1}^{\ell} + \mathbf{U}, \end{aligned}$$

respectively, where $\mathbf{Y}_0 = [\mathbf{Y}_{1,0}{}', \ldots, \mathbf{Y}_{N,0}{}']'$,

$$\begin{aligned} \mathbf{Y}_{i,0} &= [y_{i,0}, \ldots, y_{i,T}]', & \mathbf{X}_0 &= [\mathbf{X}_{1,0}', \ldots, \mathbf{X}_{N,0}']', & \mathbf{X}_{i,0} &= [\mathbf{x}_{i,0}, \ldots, \mathbf{x}_{i,T}]', \\ \mathbf{Y}^{\ell} &= \left[\mathbf{Y}_1^{\ell\prime}, \ldots, \mathbf{Y}_N^{\ell}{}'\right]', & \mathbf{Y}_i^{\ell} &= \left[y_{i,1}^{\ell}, \ldots, y_{i,T}^{\ell}\right]', & \mathbf{Y}_{-1}^{\ell} &= [\mathbf{Y}_{1,-1}^{\ell}{}', \ldots, \mathbf{Y}_{N,-1}^{\ell}{}']', \\ \mathbf{Y}_{i,-1}^{\ell} &= \left[y_{i,0}^{\ell}, \ldots, y_{i,T-1}^{\ell}\right]', & \mathbf{Y}_0^{\ell} &= \left[\mathbf{Y}_{1,0}^{\ell}{}', \ldots, \mathbf{Y}_{N,0}^{\ell}{}'\right]', & \mathbf{Y}_{i,0}^{\ell} &= \left[y_{i,0}^{\ell}, \ldots, y_{i,T}^{\ell}\right]', \end{aligned}$$

and $\mathbf{X}_0$ is assumed to have full column rank. By combining the observable and latent equations the model can equivalently be written

$$y_{i,t} = \alpha y_{i,t-1} + \mathbf{x}_{i,t}'\boldsymbol{\beta} - \mathbf{x}_{i,t-1}'\boldsymbol{\beta}\alpha + u_{i,t}, \quad t = 1, \ldots, T, \tag{2.8}$$

or, in matrix form,

$$\mathbf{Y} = \alpha\mathbf{Y}_{-1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{U}, \tag{2.9}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\beta}', -\boldsymbol{\beta}'\alpha]'$, $\mathbf{Y}_{-1} = [\mathbf{Y}_{1,-1}', \ldots, \mathbf{Y}_{N,-1}']'$,

$$\begin{aligned} \mathbf{Y}_{i,-1} &= [y_{i,0}, \ldots, y_{i,T-1}]', & \mathbf{X} &= [\mathbf{X}_1', \ldots, \mathbf{X}_N']', & \mathbf{X}_i &= [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T}]', \\ \mathbf{X}_{-1} &= [\mathbf{X}_{1,-1}', \ldots, \mathbf{X}_{N,-1}']', & \mathbf{X}_{i,-1} &= [\mathbf{x}_{i,0}, \ldots, \mathbf{x}_{i,T-1}]', & \text{and } \mathbf{Z} &= [\mathbf{X}, \mathbf{X}_{-1}]. \end{aligned}$$

We are particularly concerned with the following two special cases:

$$\sigma_i^2 = \sigma^2 \; \forall i \tag{M1}$$

$$\mathbf{X}_0 = \mathbf{I}_N \otimes \mathbf{X}_{1,0} \tag{M2}$$

where here and in the sequel, we denote by $\otimes$ the Kronecker product. Model (M1) allows for

an arbitrary set of exogenous regressors, but restricts the individual variances to be equal. In model (M2), on the other hand, the regressor matrices are assumed identical for all individuals, but there is no assumption on the individual variances.

2.4 Estimation by Least Squares

2.4.1 Model (M1): Arbitrary Regressors, Identical Variance

This section applies the method of quantile unbiased estimating functions to the least squares estimation of the autoregressive coefficient in model (M1), thus slightly generalizing the procedure of Phillips and Sul (2003), and embedding it in our methodology. From the Frisch–Waugh theorem, the least squares estimator $\hat{\alpha}_{\mathrm{LS}}$ can be expressed as

$$c : \eta_1(c) \equiv \frac{\mathbf{Y}'_{-1}\mathbf{M}\mathbf{Y}}{\mathbf{Y}'_{-1}\mathbf{M}\mathbf{Y}_{-1}} - c = 0, \tag{2.10}$$

where $\mathbf{M} = \mathbf{I}_{NT} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$. Generalizing a result of Phillips and Sul (2003), it is shown in the appendix that η_1 has distribution free of nuisance parameters, as required by Proposition 1.

2.4.1.1 Quantile Unbiased Estimation

Computation of (2.5) requires a method of evaluating the distribution function of η_1. With $\boldsymbol{\Upsilon} = \{\mathbf{Y}_0, \mathbf{X}_0\}$, $\theta = \alpha$, and $\boldsymbol{\delta} = \{\boldsymbol{\beta}, \{\sigma_i\}\}$, it is shown in the appendix that the quantile unbiased estimator solves

$$\Pr_c\left(\frac{\mathbf{U}'_0\mathbf{A}\mathbf{U}_0}{\mathbf{U}'_0\mathbf{B}\mathbf{U}_0} \leq \hat{\theta}(\boldsymbol{\Upsilon})\right) - q = 0, \tag{2.11}$$

where matrices $\mathbf{A} = \mathbf{A}(c)$ and $\mathbf{B} = \mathbf{B}(c)$ are as in (2.36), $\mathbf{U}_0 \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}_{NT+N}\right)$, and $\hat{\theta}(\boldsymbol{\Upsilon}) = \frac{\mathbf{Y}'_{-1}\mathbf{M}\mathbf{Y}}{\mathbf{Y}'_{-1}\mathbf{M}\mathbf{Y}_{-1}}$ is the observed OLS estimator.

While Andrews (1993) uses the Imhof (1961) algorithm to evaluate the requisite distribution function and Phillips and Sul (2003) resort to simulation, we replace these time-consuming processes by a saddlepoint approximation. More generally, and as will be required in the non-homoskedastic case, it is shown in the appendix that an approximation to the distribution function of

$$\bar{R} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{U}'_i\mathbf{A}_1\mathbf{U}_i}{\mathbf{U}'_i\mathbf{A}_2\mathbf{U}_i},$$

the mean of N i.i.d. ratios of quadratic forms in standard normal variates, is given by

$$\Pr\left(\bar{R} \leq \bar{r}\right) \approx \Phi\left(\hat{w}_n + \frac{1}{\hat{w}_n}\log\frac{\hat{u}_n}{\hat{w}_n}\right), \quad \bar{r} \neq \mu,$$

where

$$\hat{w}_n = \sqrt{N\log|\mathbf{D}|}\,\mathrm{sgn}(\bar{r} - \mu),$$

$$\hat{u}_n = \hat{s}\sqrt{2N \operatorname{tr} \mathbf{K}_3^2}\left[\frac{\left(2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3 + \operatorname{tr}\mathbf{K}_2\right)^2 - 4\hat{s}^2 \operatorname{tr}\mathbf{K}_2^2 \operatorname{tr}\mathbf{K}_3^2}{\operatorname{tr}^2 \mathbf{K}_2}\right]^{(N-1)/2},$$

$\mathbf{D} = \mathbf{I} - 2\hat{s}\mathbf{A}_3$, $\mathbf{A}_3 = \mathbf{A}_1 - \bar{r}\mathbf{A}_2$, $\mu = \operatorname{tr}\mathbf{A}_1 / \operatorname{tr}\mathbf{A}_2$, $\mathbf{K}_i = \mathbf{A}_i\mathbf{D}^{-1}$, $i \in \{2, 3\}$, Φ denotes the standard normal cdf, and, with λ_i denoting the eigenvalues of $\mathbf{A}_3$, the saddlepoint $\hat{s}$ solves

$$\operatorname{tr}\mathbf{K}_3 \equiv \sum_{i=1}^{T} \frac{\lambda_i}{1 - 2\hat{s}\lambda_i} = 0.$$

This generalizes the result for the $N = 1$ case as given in Lieberman (1994a) and is of interest in itself, as it is potentially applicable in numerous other modelling contexts. Also, unlike in the $N = 1$ case, no exact methods exist for evaluating the distribution function, rendering the saddlepoint approximation the only practical means of its computation.

Calculation of $\hat{\alpha}_q$ thus only requires a univariate root search over c in (2.11). Efficient estimates and approximate confidence intervals for $\boldsymbol{\beta}$ can be computed as usual from a GLS estimation using $\alpha = \hat{\alpha}_{\text{Med}}$.

2.4.1.2 Mean Adjusted Estimation

We now turn to the construction of the mean adjusted estimator of α in model (M1). It is given as the solution to

$$\eta_1(c) - \mathbb{E}_c\left[\eta_1(c)\right] = 0.$$

In order to evaluate

$$\mathbb{E}_c\big[\eta_1(c)\big] = \mathbb{E}_c\left(\frac{\mathbf{U}_0'\mathbf{A}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{B}\mathbf{U}_0}\right) - c, \tag{2.12}$$

we make use of the expression for the mean of a ratio of quadratic forms in normal variates given in Sawa (1978). Let $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ be the spectral decomposition of $\mathbf{B}$ and set $\mathbf{C} = \mathbf{P}'\mathbf{A}\mathbf{P}$. Then

$$\mathbb{E}\left[\frac{\mathbf{U}_0'\mathbf{A}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{B}\mathbf{U}_0}\right] = \int_0^\infty \sum\nolimits_{j=1}^{T+1} \frac{c_j}{(1 + 2\lambda_j t)^{3/2} \prod\limits_{k \neq j} (1 + 2\lambda_k t)^{1/2}} \mathrm{d}t, \tag{2.13}$$

where c_j and λ_j denote the j^{th} diagonal element of $\mathbf{C}$ and $\boldsymbol{\Lambda}$, respectively. The integrand in (2.13) dies off quickly, so that the integral is straightforward to evaluate numerically.

This method of bias correction was employed by Tanizaki (2000) in the pure time–series case, and by Gouriéroux et al. (2006) in the DPD setting, both of whom used simulation for the evaluation of the mean function.

In view of Equation (2.6), it is apparent that $\mathbb{E}_c\left[\eta_1(c)\right] \equiv b_{\hat{\alpha}_{\text{LS}}}(c)$ is just the bias of the least squares estimator under a true parameter of c. If the model includes only individual dummies,

this bias is asymptotically (Nickell, 1981; Hahn and Kuersteiner, 2002):

$$\lim_{N\to\infty} b_{\hat{\alpha}_{\mathrm{LS}}}(c) = \begin{cases} -\frac{1+c}{T-1}\left(1-\frac{1-c^T}{T(1-c)}\right)\left(1-\frac{2c}{(1-c)(T-1)}\left(1-\frac{1-c^T}{T(1-c)}\right)\right)^{-1}, & \text{if } \frac{N}{T}\to\infty, \\ -\frac{1+c}{T}, \quad \text{if } \frac{N}{T}\to k<\infty,\ \frac{N}{T^3}\to 0. \end{cases}$$

Expressions for the asymptotic bias in the more general model with exogenous regressors are also available (Kiviet, 1995; Bun and Kiviet, 2003; Phillips and Sul, 2007). Bruno (2005) extends these to unbalanced panels.

Replacing $\mathbb{E}_c[\eta_1(c)]$ in (2.12) with the above limit as $N/T \to \infty$ and $N/T \to k < \infty$, we obtain the estimators of Bun and Carree (2005) and Hahn and Kuersteiner (2002), respectively. As such, our estimator can be seen as a version of these estimators valid when both N and T are small, and for arbitrary sets of exogenous regressors.

2.4.2 Model (M2): Identical Regressors, Arbitrary Variance

In this section we relax the assumption of homoskedasticity under some additional constraints. In particular, we drop the assumption that $\sigma_i^2 = \sigma^2$, so that now, for $i = 1, \dots, N$, the errors of the individual series satisfy $u_{it} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_i^2)$, $t = 1, \dots, T$.

The regressor matrix is assumed to be block diagonal, so that

$$\mathbf{X}_0 = I_N \otimes \mathbf{X}_{1,0},$$

where $\mathbf{X}_{1,0}$ represents the $(T+1) \times k_1$ individual regressor matrix. The coefficient vector $\boldsymbol{\beta}$ can be partitioned as $(\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_N')'$, where each of the $\boldsymbol{\beta}_i$ has $k_1 \equiv k/N$ elements. This restricted model encompasses the standard fixed effect model which includes a dummy regressor for each individual, and the models considered in Phillips and Sul (2003).

Let $\mathbf{V}_u = \mathrm{diag}\left(\sigma_1^2, \dots, \sigma_N^2\right) \otimes I_T$. The model in matrix form is then given by

$$\begin{aligned} \mathbf{Y}_0 &= \mathbf{X}_0\boldsymbol{\beta} + \mathbf{Y}_0^{\ell}, \\ \mathbf{Y}^{\ell} &= \alpha\mathbf{Y}_{-1}^{\ell} + \mathbf{U}, \qquad \mathbf{U} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_u), \end{aligned} \tag{2.14}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_N')'$, or, equivalently,

$$\mathbf{Y} = \alpha\mathbf{Y}_{-1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{U}, \qquad \mathbf{U} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_u), \tag{2.15}$$

where now $\mathbf{Z} = \mathbf{I}_N \otimes \mathbf{Z}_1$, $\mathbf{Z}_1 = [\mathbf{X}_1 \quad \mathbf{X}_{1,-1}]$, $\mathbf{X}_1$ and $\mathbf{X}_{1,-1}$ are defined analogously to $\mathbf{X}_{1,0}$ by omitting the observations at time $t = 0$ and $t = T$, respectively, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \dots, \boldsymbol{\gamma}_N')'$ and $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i', -\alpha\boldsymbol{\beta}_i')'$. If $\mathbf{Z}_1$ is singular, then it should be replaced by a full-rank matrix spanning the same column space.

Applying the Frisch-Waugh theorem to (2.15) premultiplied by $\mathbf{V}_u^{-1/2} = \mathrm{diag}\left(\sigma_1^{-1}, \dots, \sigma_N^{-1}\right)$

$\otimes I_T$, we arrive at the GLS estimator $\hat{\alpha}_{\text{GLS}}$, which solves

$$c : \frac{\mathbf{Y}'_{-1}\mathbf{V}_u^{-1/2}\mathbf{M}\mathbf{V}_u^{-1/2}\mathbf{Y}}{\mathbf{Y}'_{-1}\mathbf{V}_u^{-1/2}\mathbf{M}\mathbf{V}_u^{-1/2}\mathbf{Y}_{-1}} - c = 0,$$

where

$$\begin{aligned}\mathbf{M} &= I_{NT} - \mathbf{V}_u^{-1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}_u^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}_u^{-1/2} \\ &= I_{NT} - I_N \otimes \mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1' \equiv \mathbf{I}_N \otimes \mathbf{M}_1.\end{aligned}$$

Due to the simple structure of matrices $\mathbf{M}$ and $\mathbf{V}_u$, $\hat{\alpha}_{\text{GLS}}$ can be equivalently written as the solution of

$$c : \frac{\sum_{i=1}^N \frac{1}{\sigma_i^2}\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_i}{\sum_{i=1}^N \frac{1}{\sigma_i^2}\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_{-1,i}} - c = 0.$$

Upon estimating the individual variances, for given c, by

$$\hat{\sigma}_i^2 = \frac{\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_{i,-1}}{\operatorname{tr}\mathbf{R}_1'\mathbf{M}_1\mathbf{R}_1},$$

where $\mathbf{R}_1 = \mathbf{R}_1(c)$ is as in (2.35), we obtain the feasible GLS estimator

$$c : \eta_2(c) \equiv \frac{1}{N}\sum_{i=1}^N \frac{\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_i}{\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_{i,-1}} - c = 0. \tag{2.16}$$

As $\eta_2(c)$ is the average of the individual OLS estimating functions, its independence of $\boldsymbol{\beta}$ and the σ_i follows from the corresponding property of the OLS estimating function proven in the appendix.

2.4.2.1 Quantile Unbiased Estimation

Following through the same steps that led to (2.36) shows that

$$\Pr_c\left(\eta_2(c) \le \eta_2(c, \boldsymbol{\Upsilon})\right) = \Pr_c\left(\frac{1}{N}\sum_{i=1}^N \frac{\mathbf{U}'_{i,0}\mathbf{A}_1\mathbf{U}_{i,0}}{\mathbf{U}'_{i,0}\mathbf{B}_1\mathbf{U}_{i,0}} \le \hat{\theta}(\boldsymbol{\Upsilon})\right),$$

where $\mathbf{A}_1(c) = \frac{1}{2}\left[\mathbf{A}_1^{*\prime} + \mathbf{A}_1^*\right]$, $\mathbf{A}_1^* = \mathbf{R_1}'\mathbf{D}'_{T-1}\mathbf{M}_1\mathbf{D}_T\mathbf{R}_1$, $\mathbf{B}_1(c) = \mathbf{R}_1'\mathbf{D}'_{T-1}\mathbf{M}_1\mathbf{D}_{T-1}\mathbf{R}_1$, $\mathbf{U}_{i,0} \overset{\text{iid}}{\sim} \text{N}\left(\mathbf{0}, \mathbf{I}_{N(T+1)}\right)$, and $\hat{\theta}(\boldsymbol{\Upsilon}) = \frac{1}{N}\sum_{i=1}^N \frac{\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_i}{\mathbf{Y}'_{i,-1}\mathbf{M}_1\mathbf{Y}_{i,-1}}$ is the observed feasible GLS estimator. This can be evaluated by means of the saddlepoint approximation for the mean of N i.i.d. ratios of quadratic forms as discussed in Section 2.4.1.1.

2.4.2.2 Mean Adjusted Estimation

The mean unbiased estimating equation is given by

$$\eta_2(c) - \mathbb{E}_c\left[\eta_2(c)\right] = 0.$$

It is immediate that

$$\mathbb{E}_c\big[\eta_2(c)\big] = \mathbb{E}_c\left[\frac{\mathbf{U}_{1,0}'\mathbf{A}_1\mathbf{U}_{1,0}}{\mathbf{U}_{1,0}'\mathbf{B}_1\mathbf{U}_{1,0}}\right] - c,$$

which can be evaluated by (2.13) with matrices $\mathbf{A}_1$ and $\mathbf{B}_1$ replacing $\mathbf{A}$ and $\mathbf{B}$, respectively.

2.5 Estimation by Maximum Likelihood

This section is concerned with the estimation of the model by maximum likelihood methods, in both the homoskedastic and heteroskedastic settings. The general idea is to concentrate out all nuisance parameters from the log likelihood, giving rise to the profile log likelihood and profile score functions. The latter takes the form of a sum of ratios of quadratic forms, to which our methodology can be applied.

2.5.1 Model (M1): Arbitrary Regressors, Identical Variance

The log likelihood of model (M1), after dropping the constant, is given by

$$\ell(\alpha, \boldsymbol{\beta}, \sigma^2) = -\frac{N(T+1)}{2}\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\sigma^2}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right),$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\alpha) = \mathbf{R}\mathbf{R}' = \mathbf{I}_N \otimes \mathbf{R}_1\mathbf{R}_1'$ with $\mathbf{R}$ and $\mathbf{R}_1 = \mathbf{R}_1(\alpha)$ as in (2.35). The score functions are

$$\dot{\ell}_{\boldsymbol{\beta}}(\alpha, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}\mathbf{X}_0'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right),$$

$$\dot{\ell}_{\sigma^2}(\alpha, \boldsymbol{\beta}, \sigma^2) = -\frac{N(T+1)}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right),$$

and

$$\dot{\ell}_{\alpha}(\alpha, \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_\alpha\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_\alpha\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right),$$

where $\dot{\mathbf{\Sigma}}_\alpha$ denotes the elementwise derivative of $\mathbf{\Sigma}$ with respect to α, given by $\dot{\mathbf{\Sigma}}_\alpha = \mathbf{I}_N \otimes \left(\mathbf{R}_1\dot{\mathbf{R}}_1' + \dot{\mathbf{R}}_1\mathbf{R}_1'\right)$, where

$$\dot{\mathbf{R}}_1 = \begin{bmatrix} b' & 0 & 0 & \cdots & 0 & 0 \\ b'\alpha + b & 0 & 0 & \cdots & 0 & 0 \\ b'\alpha^2 + 2b\alpha & 1 & 0 & \cdots & 0 & 0 \\ b'\alpha^3 + 3b\alpha^2 & 2\alpha & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b'\alpha^T + Tb\alpha^{T-1} & (T-1)\alpha^{T-2} & (T-2)\alpha^{T-3} & \cdots & 1 & 0 \end{bmatrix},$$

$b' = \alpha\left(1-\alpha^2\right)^{-3/2}$ if $\alpha \in (-1,1)$ and zero if $\alpha = 1$. It is computationally advantageous to note that $\mathbf{\Sigma}^{-1} = \mathbf{I}_N \otimes \mathbf{R}_1^{-T}\mathbf{R}_1^{-1}$, and $\mathbf{R}_1^{-1}$ is a matrix that has top left element b^{-1}, ones on the rest of main diagonal, $-\alpha$ on the first subdiagonal, and zeroes everywhere else.

Solving for $\boldsymbol{\beta}$ and σ^2, we obtain

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(\mathbf{X}_0'\mathbf{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{\Sigma}^{-1}\mathbf{Y}_0$$

and

$$\hat{\sigma}^2(\alpha) = \frac{1}{N(T+1)}\left(\mathbf{Y}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}(\alpha)\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}(\alpha)\right),$$

which can be substituted back into $\dot{\ell}_\alpha$ to yield the profile score function

$$\begin{aligned}\dot{\ell}_\alpha(\alpha,\boldsymbol{\beta},\sigma^2)\Big|_{\hat{\boldsymbol{\beta}}(\alpha),\hat{\sigma}^2(\alpha)} &= -\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\dot{\mathbf{\Sigma}}_\alpha\right) + \frac{N(T+1)}{2}\frac{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\dot{\mathbf{\Sigma}}_\alpha\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0}{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0} \\ &= -N\operatorname{tr}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right) + \frac{N(T+1)}{2}\frac{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\dot{\mathbf{\Sigma}}_\alpha\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0}{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0},\end{aligned}$$

with idempotent matrix $\mathbf{M}_{\mathbf{\Sigma}} \equiv \mathbf{I}_{N(T+1)} - \mathbf{X}_0\left(\mathbf{X}_0'\mathbf{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{\Sigma}^{-1}$.

2.5.1.1 Quantile Unbiased Estimation

Denote by $\eta_3(\alpha)$ the stochastic part of the profile score function, i.e., let

$$\eta_3(c) \equiv \frac{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\dot{\mathbf{\Sigma}}_c\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0}{\mathbf{Y}_0'\mathbf{M}_{\mathbf{\Sigma}}'\mathbf{\Sigma}^{-1}\mathbf{M}_{\mathbf{\Sigma}}\mathbf{Y}_0}, \tag{2.17}$$

where for the sake of notational continuity we have changed the argument of η_3 (and the matrices involved) from α to c. We require an expression for the probability

$$\Pr_c\left(\eta_3(c) \leq \eta_3(c,\boldsymbol{\Upsilon})\right),$$

where $\eta_3(c, \boldsymbol{\Upsilon})$ is the observed value of the estimating function at c.

It is clear from scaling and the fact that $\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{X}_0 = \mathbf{0}$ that $\eta_3(c)$ has distribution free of $(\boldsymbol{\beta}, \sigma^2)$, so that

$$\begin{aligned}
\Pr_c\left(\eta_3(c) \leq \eta_3(c, \boldsymbol{\Upsilon})\right) &= \Pr_c\left(\frac{\left(\mathbf{Y}_0^{\ell}\right)' \mathbf{M}'_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}_0^{\ell}}{\left(\mathbf{Y}_0^{\ell}\right)' \mathbf{M}'_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}_0^{\ell}} \leq \eta_3(c, \boldsymbol{\Upsilon})\right) \\
&= \Pr\left(\frac{\mathbf{U}_0'\mathbf{R}'\mathbf{M}'_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{R}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{R}'\mathbf{M}'_{\boldsymbol{\Sigma}}\mathbf{R}^{-T}\mathbf{R}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{R}\mathbf{U}_0} \leq \eta_3(c, \boldsymbol{\Upsilon})\right) \\
&= \Pr\left(\frac{\mathbf{U}_0'\mathbf{M}_{\mathbf{R}}\mathbf{R}^{-1}\dot{\boldsymbol{\Sigma}}_c\mathbf{R}^{-T}\mathbf{M}_{\mathbf{R}}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{M}_{\mathbf{R}}\mathbf{U}_0} \leq \eta_3(c, \boldsymbol{\Upsilon})\right),
\end{aligned}$$

where the last equality follows with

$$\begin{aligned}
\mathbf{R}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{R} &= \mathbf{R}^{-1}(\mathbf{I}_{N(T+1)} - \mathbf{X}_0\left(\mathbf{X}_0'\boldsymbol{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\boldsymbol{\Sigma}^{-1})\mathbf{R} \\
&= \mathbf{R}^{-1}(\mathbf{I}_{N(T+1)} - \mathbf{X}_0\left(\mathbf{X}_0'\boldsymbol{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{R}^{-T}\mathbf{R}^{-1})\mathbf{R} \\
&= \mathbf{I}_{N(T+1)} - \mathbf{R}^{-1}\mathbf{X}_0\left(\mathbf{X}_0'\boldsymbol{\Sigma}^{-1}\mathbf{X}_0\right)^{-1}\mathbf{X}_0'\mathbf{R}^{-T} \equiv \mathbf{M}_{\mathbf{R}},
\end{aligned}$$

with symmetric and idempotent matrix $\mathbf{M}_{\mathbf{R}}$ so defined. Therefore, the quantile unbiased estimating equation becomes

$$c : \Pr\left(\frac{\mathbf{U}_0'\mathbf{A}(c)\mathbf{U}_0}{\mathbf{U}_0'\mathbf{B}(c)\mathbf{U}_0} \leq \eta_3(c, \boldsymbol{\Upsilon})\right) - q = 0, \tag{2.18}$$

where $\mathbf{A}(c) = \mathbf{M}_{\mathbf{R}}\mathbf{R}^{-1}\dot{\boldsymbol{\Sigma}}_c\mathbf{R}^{-T}\mathbf{M}_{\mathbf{R}}$ and $\mathbf{B}(c) = \mathbf{M}_{\mathbf{R}}$. This lends itself to calculation by means of our proposed saddlepoint approximation.

2.5.1.2 Mean Adjusted Estimation

A mean unbiased estimating function can be constructed by subtracting from (2.17) its mean under a true parameter of c, given by

$$\begin{aligned}
\mathbb{E}_c\left[\eta_3(c)\right] &= \mathbb{E}_c\left[\frac{\mathbf{Y}_0'\mathbf{M}'_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}_0}{\mathbf{Y}_0'\mathbf{M}'_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\boldsymbol{\Sigma}}\mathbf{Y}_0}\right] = \mathbb{E}_c\left[\frac{\mathbf{U}_0'\mathbf{M}_{\mathbf{R}}\mathbf{R}^{-1}\dot{\boldsymbol{\Sigma}}_c\mathbf{R}^{-T}\mathbf{M}_{\mathbf{R}}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{M}_{\mathbf{R}}\mathbf{U}_0}\right] \\
&= \frac{\operatorname{tr}\mathbf{M}_{\mathbf{R}}\mathbf{R}^{-1}\dot{\boldsymbol{\Sigma}}_c\mathbf{R}^{-T}\mathbf{M}_{\mathbf{R}}}{\operatorname{tr}\mathbf{M}_{\mathbf{R}}} = \frac{2\operatorname{tr}\left(\mathbf{I}_N \otimes \mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)\mathbf{M}_{\mathbf{R}}}{N(T+1)-k},
\end{aligned} \tag{2.19}$$

where the last line follows because, due to its special form, the ratio of quadratic forms appearing in the profile score is independent of its own denominator (Pitman, 1937), whence the mean of the ratio equals the ratio of the means. This bears the advantage that unlike in the least squares based counterpart (2.13), there are no unsolved integrals in (2.19), allowing for massive savings

in computational time. The mean adjusted estimator therefore solves

$$c : \eta_3(c) - \frac{\operatorname{tr}\left(\mathbf{I}_N \otimes \mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)\mathbf{M}_{\mathbf{R}}}{N(T+1)-k} = 0. \tag{2.20}$$

This is the adjusted profile likelihood estimator of McCullagh and Tibshirani (1990), which, in this case, coincides with the marginal likelihood estimator of Wilson (1989). However, McCullagh and Tibshirani derive (2.20) only as an approximation, overlooking the fact that the mean of the ratio equals the ratio of the means. If there are no exogenous regressors, i.e., $\mathbf{M}_{\mathbf{R}} = \mathbf{I}_{N(T+1)}$ and $k = 0$, then the so-defined estimator coincides with the MLE of α.

2.5.2 Model (M2): Identical regressors, Arbitrary Variance

For the model with unequal variances and identical regressors, the log likelihood becomes, after dropping the constant,

$$\begin{aligned}
\ell\big(\alpha, \boldsymbol{\beta}, \{\sigma_i^2\}\big) &= -\frac{1}{2}\log|\mathbf{V}_u \otimes \boldsymbol{\Sigma}_1| - \frac{1}{2}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right)'\left(\mathbf{V}_u \otimes \boldsymbol{\Sigma}_1\right)^{-1}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right) \\
&= -\frac{T+1}{2}\log|\mathbf{V}_u| - \frac{N}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right)'\left(\mathbf{V}_u^{-1} \otimes \boldsymbol{\Sigma}_1^{-1}\right)\left(\mathbf{Y}_0 - \mathbf{X}_0\boldsymbol{\beta}\right) \\
&= -\frac{T+1}{2}\sum_{i=1}^{N}\log\sigma_i^2 - \frac{N}{2}\log|\boldsymbol{\Sigma}_1| - \sum_{i=1}^{N}\frac{1}{2\sigma_i^2}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right)'\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right),
\end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_1(\alpha) \equiv \mathbf{R}_1\mathbf{R}_1'$, and $\mathbf{R}_1$ is as in (2.35). The score functions are

$$\dot{\ell}_{\boldsymbol{\beta}_i}\left(\alpha, \boldsymbol{\beta}_i, \sigma_i^2\right) = \frac{1}{\sigma_i^2}\mathbf{X}_{i,0}'\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right),$$

$$\dot{\ell}_{\sigma_i^2}\left(\alpha, \boldsymbol{\beta}_i, \sigma_i^2\right) = -\frac{(T+1)}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right)'\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right),$$

and

$$\dot{\ell}_{\alpha}\left(\alpha, \boldsymbol{\beta}, \{\sigma_i^2\}\right) = -N\operatorname{tr}\left(\dot{\mathbf{R}}_1\mathbf{R}_1^{-1}\right) + \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right)'\left(\boldsymbol{\Sigma}_1^{-1}\dot{\mathbf{R}}_1\mathbf{R}_1^{-1}\right)\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\boldsymbol{\beta}_i\right),$$

from which

$$\hat{\boldsymbol{\beta}}_i(\alpha) = \left(\mathbf{X}_{i,0}'\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_{i,0}\right)^{-1}\mathbf{X}_{i,0}'\boldsymbol{\Sigma}_1^{-1}\mathbf{Y}_{i,0}$$

and

$$\hat{\sigma}_i^2(\alpha) = \frac{1}{T+1}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\hat{\boldsymbol{\beta}}_i(\alpha)\right)'\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{Y}_{i,0} - \mathbf{X}_{i,0}\hat{\boldsymbol{\beta}}_i(\alpha)\right).$$

Substituting back into $\dot{\ell}_\alpha$, we obtain the profile score function

$$\dot{\ell}_\alpha(\alpha, \boldsymbol{\beta}, \{\sigma_i^2\})\Big|_{\hat{\boldsymbol{\beta}}(\alpha), \hat{\sigma}^2(\alpha)} = -N \operatorname{tr}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right) + (T+1)\sum_{i=1}^{N} \frac{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\left(\boldsymbol{\Sigma}_1^{-1}\dot{\mathbf{R}}_1\mathbf{R}_1^{-1}\right)\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}}{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}},$$

with idempotent matrix $\mathbf{M}_{\boldsymbol{\Sigma}_1} = \mathbf{I}_{T+1} - \mathbf{X}_{i,0}\left(\mathbf{X}_{i,0}'\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_{i,0}\right)^{-1}\mathbf{X}_{i,0}\boldsymbol{\Sigma}_1^{-1}$.

2.5.2.1 Quantile Unbiased Estimation

Let

$$\eta_4(c) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\left(\boldsymbol{\Sigma}_1^{-1}\dot{\mathbf{R}}_1\mathbf{R}_1^{-1}\right)\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}}{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}} \tag{2.21}$$

denote the stochastic part of the profile score function. From scaling and the fact that $\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{X}_{i,0} = \mathbf{0}$, $\eta_4(c)$ has distribution free of $\left(\boldsymbol{\beta}, \{\sigma_i^2\}\right)$, and we have that

$$\begin{aligned}\Pr\left(\eta_4(c) \leq \eta_4(c, \boldsymbol{\Upsilon})\right) &= \Pr\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\left(\boldsymbol{\Sigma}_1^{-1}\dot{\mathbf{R}}_1\mathbf{R}_1^{-1}\right)\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}}{\mathbf{Y}_{i,0}'\mathbf{M}_{\boldsymbol{\Sigma}_1}'\boldsymbol{\Sigma}_1^{-1}\mathbf{M}_{\boldsymbol{\Sigma}_1}\mathbf{Y}_{i,0}} \leq \eta_4(c, \boldsymbol{\Upsilon})\right) \\ &= \Pr\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{U}_{i,0}'\mathbf{M}_{\mathbf{R}_1}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)'\mathbf{M}_{\mathbf{R}_1}\mathbf{U}_{i,0}}{\mathbf{U}_{i,0}'\mathbf{M}_{\mathbf{R}_1}\mathbf{U}_{i,0}} \leq \eta_4(c, \boldsymbol{\Upsilon})\right),\end{aligned}$$

with symmetric and idempotent matrix

$$\mathbf{M}_{\mathbf{R}_1} = \mathbf{I}_{T+1} - \mathbf{R}_1^{-1}\mathbf{X}_{i,0}\left(\mathbf{X}_{i,0}'\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_{i,0}\right)^{-1}\mathbf{X}_{i,0}'\mathbf{R}_1^{-T}.$$

Hence, the quantile unbiased estimator solves

$$c : \Pr\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{U}_{i,0}'\mathbf{A}(c)\mathbf{U}_{i,0}}{\mathbf{U}_{i,0}'\mathbf{B}(c)\mathbf{U}_{i,0}} \leq \eta_4(c, \boldsymbol{\Upsilon})\right) - q = 0, \tag{2.22}$$

where $\mathbf{A}(c) = \mathbf{M}_{\mathbf{R}_1}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)'\mathbf{M}_{\mathbf{R}_1}$ and $\mathbf{B}(c) = \mathbf{M}_{\mathbf{R}_1}$.

2.5.2.2 Mean Adjusted Estimation

Similar to Section 2.5.1.2, the mean of $\eta_4(c)$ in (2.21) under a true parameter of c is

$$\mathbb{E}_c\left[\eta_4(c)\right] = \frac{\operatorname{tr}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)'\mathbf{M}_{\mathbf{R}_1}}{T+1-k_1},$$

so that the mean unbiased estimating equation can be written

$$c : \eta_4(c, \boldsymbol{\Upsilon}) - \frac{\operatorname{tr}\left(\mathbf{R}_1^{-1}\dot{\mathbf{R}}_1\right)' \mathbf{M}_{\mathbf{R}_1}}{T+1-k_1} = 0.$$

This is a panel version of the adjusted profile likelihood estimator of McCullagh and Tibshirani (1990), and also the marginal likelihood estimator of Wilson (1989). As before, if there are no exogenous regressors, then the estimator coincides with the MLE of α.

2.6 Numerical Results

2.6.1 Point Estimation

This section describes the results of a simulation study we conducted in order to exemplify the virtues of our proposed point estimators, for both models (M1) and (M2). The model under investigation includes individual dummies as the only exogenous variables, so as to facilitate comparison with other estimators that are specifically designed for that setup; a different set of regressors would require extensive custom calculations. The sample sizes we considered include all possible combinations of $N \in \{10, 25, 50\}$ and $T+1 \in \{10, 25, 50\}$, for each of which $1,000$ samples of innovations were drawn from the standard normal distribution. The only difference between the two data generating processes is that for model (M2), we multiplied the innovation sequence for one of the individuals by N, so that $\sigma_N = N$. This mimics a panel in which one of the individuals is "large" compared to the remaining ones, as occurs, e.g., in macroeconomic panels that include the US. For all sample sizes under investigation, the relative performances of the estimators were qualitatively similar; as such, we only report the results for $N = 10$ and $T + 1 = 10$, a somewhat typical setup when working with macroeconomic data. The results for other configurations are available from the authors upon request.

The top row of Figure 2.1 shows the results for model (M1). In an attempt to unclutter the graphs, we do not show the results for the mean adjusted estimator and the OLS based estimators: their performance was generally very close to that of the quantile unbiased ML-based estimator (2.18), represented by solid lines. Not surprisingly, in this small-N, small-T setup, both the OLS estimator and the GMM estimator of Arellano and Bond (1991), shown as dashes and dots respectively, exhibit substantial bias. This is in stark contrast to the quantile unbiased estimator, which is slightly mean biased only near $\alpha = 1$, and virtually median unbiased. This naturally has dramatic consequences for the root mean squared error (RMSE), shown in the rightmost column. In terms of this latter measure, the quantile unbiased estimator improves on the RMSE of the standard estimators by a factor of up to 6. The asymptotically unbiased estimators of Hahn and Kuersteiner (2002) (crosses) and Bun and Carree (2005) (dash-dot) perform between the two extremes, being outperformed by the quantile unbiased estimator only near the stationarity border. It should, however, be borne in mind that these letter two estimators are specifically

tailored to this very model, whereas the other estimator can accommodate any set of exogenous regressors.

The bottom row of the same figure depicts the results for model (M2). As the theory suggests, the ML-based median unbiased estimator (2.22) performs favorably in terms of both bias and RMSE.[1] It is, in fact, *invariant* with respect to the individual variances, whereas the other estimators take a dramatic impact from a departure from cross-sectional homogeneity such as in the data generating process used in this study. Of course those estimators were not designed for that model, and in fact had we chosen estimator (2.18) instead of the one designed for model (M2), it would have suffered the same fate; however, a comparison of the top and bottom rows of Figure 2.1 shows that the ML-based median unbiased estimator suffers almost no loss in accuracy as the assumption of equal individual variances is dropped. As such, the need for that unrealistic assumption should hardly ever arise.

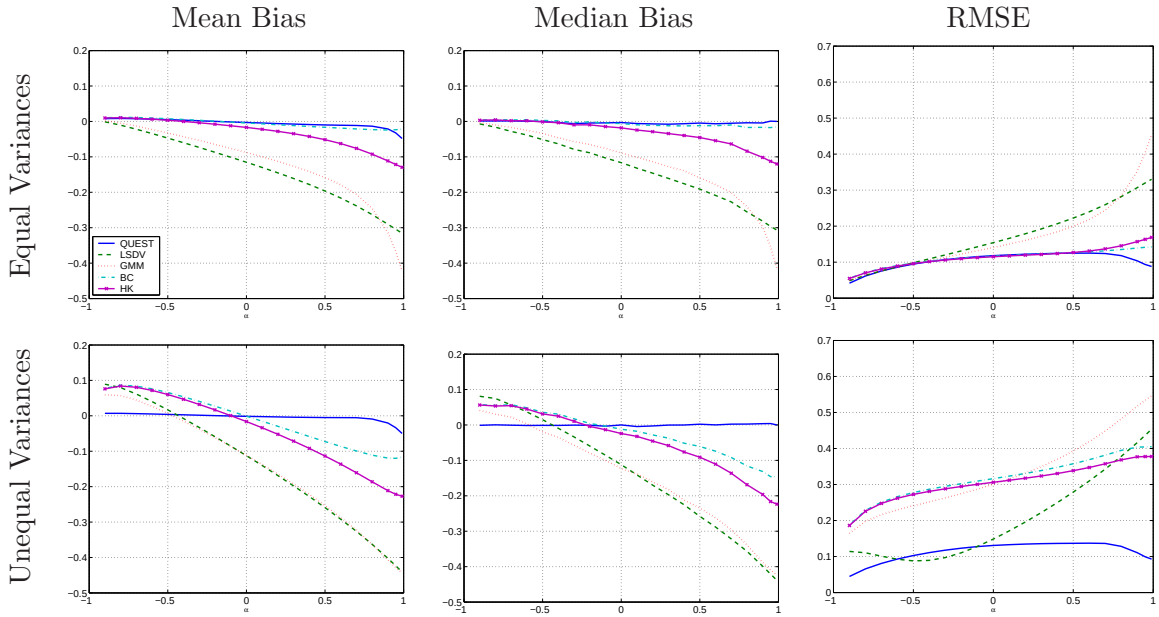

Figure 2.1: Bias and Root Mean Squared Error of different estimators of the autoregressive coefficient, for a model with $N = 10$, $T + 1 = 10$, $\mathbf{x}_{it} = [1]$, $\sigma_i = 1\ \forall i \in \{1, \ldots, N-1\}$. The top and bottom rows use $\sigma_N = 1$ and $\sigma_N = N$, respectively.

2.6.2 Interval Estimation

This section investigates the quality of the ML and OLS–based quantile unbiased interval estimates. For all combinations of $N \in \{10, 25, 50\}$ and $T + 1 \in \{10, 25, 50\}$, and using the same

[1] The (not shown) OLS–based median-"unbiased" estimator shows a somewhat different picture: while still being far ahead of the GMM OLS estimators, it is no longer (approximately exactly) median unbiased, especially for values of α close to the stationarity border; this must be attributed to the loss of accuracy of the associated saddlepoint approximation, as explained in the appendix. The OLS–based mean adjusted estimator, on the other hand, does not suffer from this deficiency, and still performs close to the ML-based variant.

samples as in Section 2.6.1, 1,000 90% equal-tail intervals for α were computed based on (2.5) with $q = (1 \pm 0.9)/2$. Figure 2.2 contains the results for the homoskedastic case, with the average length of the intervals being shown on the left scale, and the empirical coverage on the right scale. The dotted lines represent 5% critical values of a two-sided binomial test of the null hypothesis that the true coverage of each intervals equals its nominal value, i.e., 90%. Similar to the results for the point estimators in the homoskedastic case, the results for the OLS (dashes) and MLE (solid) based intervals do not differ much, with empirical coverages lying well within the critical values for all sample sizes shown.

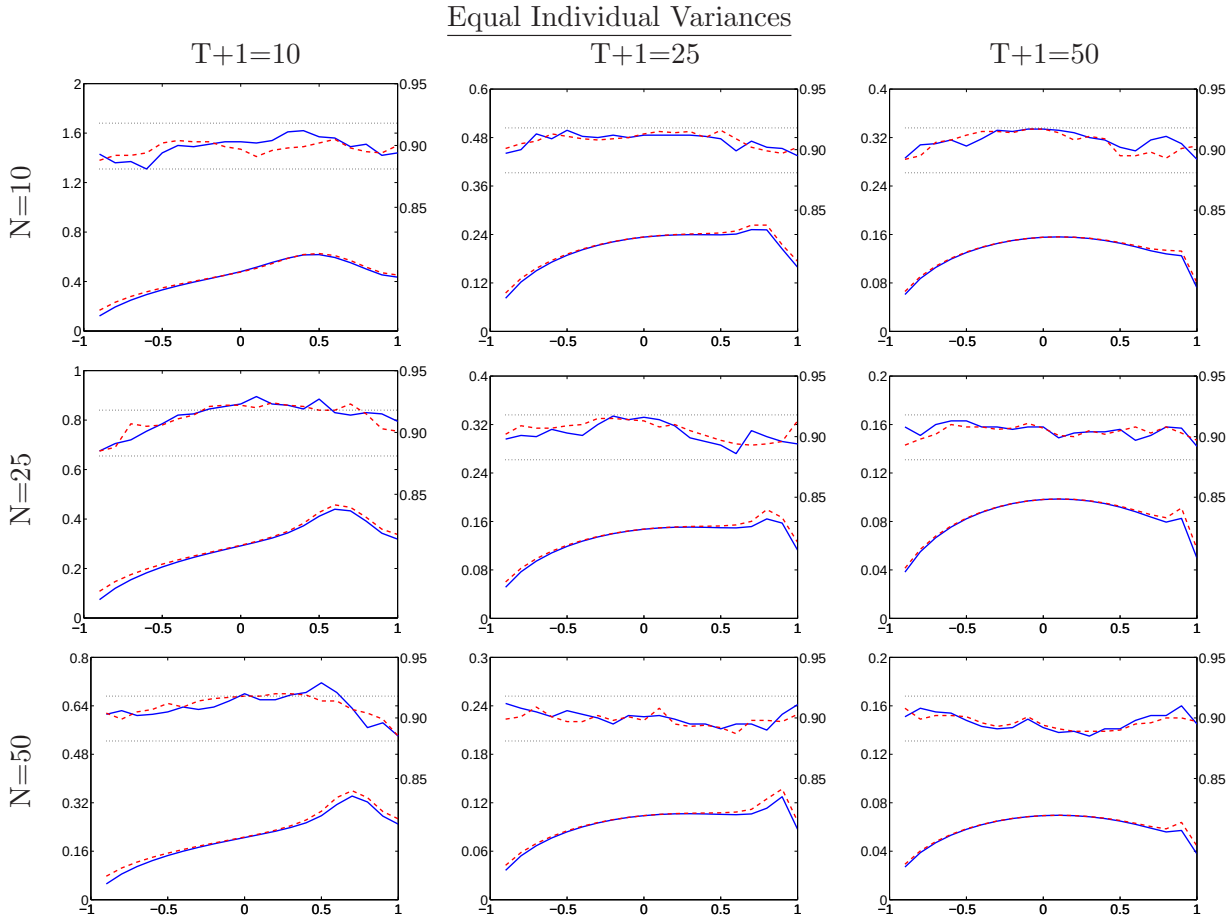

Figure 2.2: Mean length (left scale) and empirical coverage (right scale) of OLS–based (dashes) and ML-based (solid) $\delta = 90\%$ equal tails confidence intervals. The dotted lines represent 5% critical values of a two sided test of $H_0 : \delta = 0.9$.

The differences between estimators are substantially more pronounced for the heteroskedastic case. Figure 2.3 contains the results. While for the ML-based confidence intervals, the null hypothesis that the true coverage equals the nominal value is again well supported, this is not true for the OLS–based intervals, even though the latter tend to be considerably longer on average. As before, this must be attributed to the decreased accuracy of the associated saddlepoint

approximation. In particular, the accuracy of the approximation deteriorates as i) N increases, ii) T decreases, and iii) α approaches the stationarity border. This is exactly reflected in the performance of the OLS–based interval estimates.

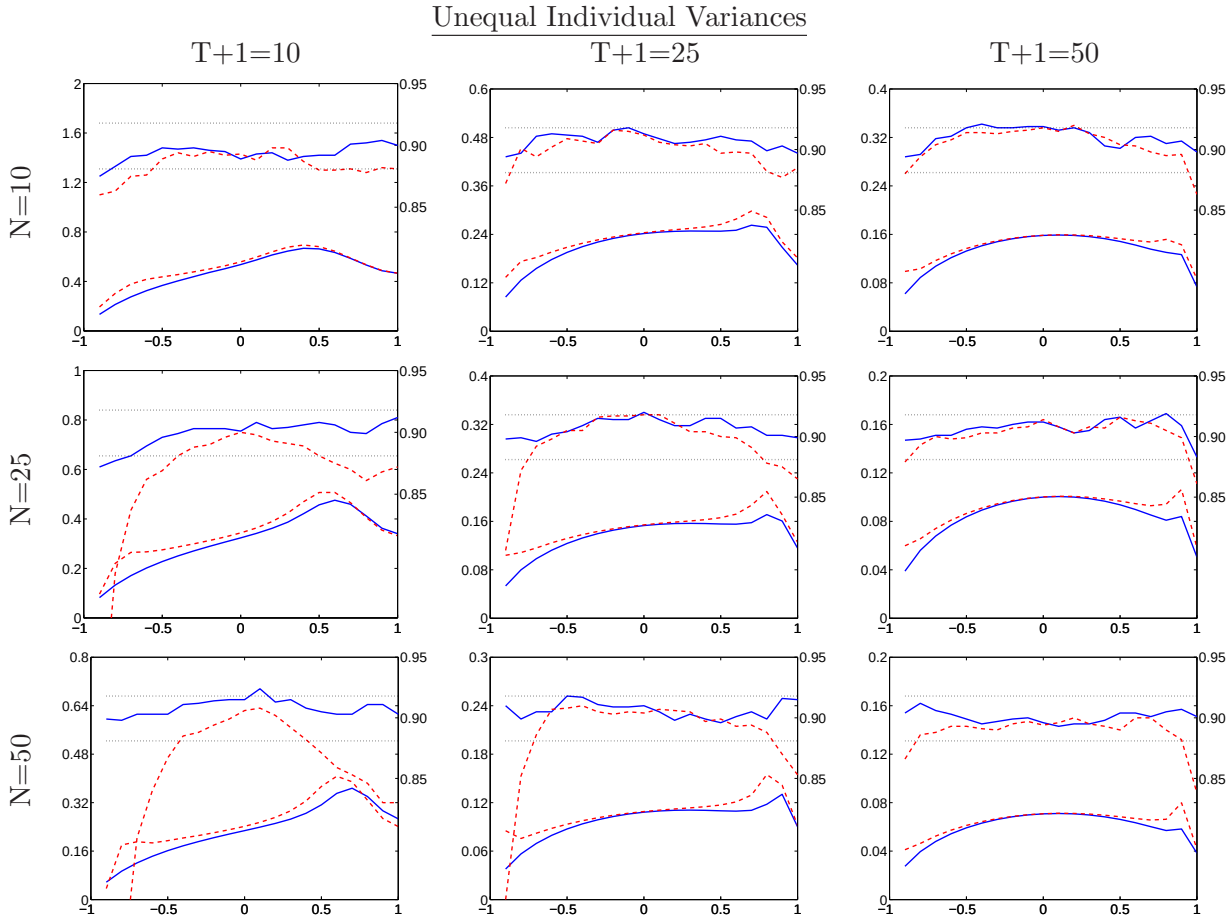

Figure 2.3: Mean length (left scale) and empirical coverage (right scale) of OLS–based (dashes) and ML-based (solid) $\tau = 90\%$ equal tails confidence intervals. The dotted lines represent 5% critical values of a two sided test of $H_0 : \tau = 0.9$.

The comments made in the previous section regarding the relative performance of the ML-based estimators in the homoskedastic and heteroskedastic cases remain valid in the context of interval estimation. Considering that in many, if not most, empirical applications, the assumption of equal individual variances is questionable at best, this behavior is fortuitous, as it will allow researchers to abandon this restriction altogether.

2.7 Conclusions

Most bias corrected estimators for dynamic panel models that have been considered in the literature rely on either simulation or asymptotics to obtain the required bias functions. Both

approaches have their limitations: the former in terms of computational requirements, the latter in terms of generality. Moreover, most of the previously proposed methods impose the unrealistic assumption of variance homogeneity, and are incapable of delivering reliable interval estimates.

The concepts considered in the present manuscript overcome these limitations. The two cornerstones of our approach are the notion of a quantile unbiased estimating function, and a saddlepoint approximation to the tail probabilities of the original estimating function. The method is capable of extremely fast and accurate inference in the first–order dynamic panel model, without having to rely on any asymptotic arguments. As regards point estimation, for example, not only does it successfully tackle the problem of asymptotic bias, but essentially *removes* the bias, even for samples as small as $N = 10$, $T + 1 = 10$. We consider it an important feature of our likelihood–based method that the properties of both the point and interval estimates are virtually unaltered as the assumption of equal individual variances is dropped, as it allows us to recommend that this restriction be abandoned wherever it is maintained for the sole sake of feasibility.

The methods developed herein are not restricted to the first–order DPD model. The concept of quantile unbiased estimation equations is quite general, and especially useful in linear models with a covariance matrix depending on one parameter. The saddlepoint approximation to the mean of ratios of quadratic forms has other applications as well, certainly in panel model contexts. We leave those for future research.

Appendix 2.A A Saddlepoint Approximation for the Mean of i.i.d. Ratios of Quadratic Forms

This section develops saddlepoint approximations for the density and distribution function of the mean of N i.i.d. quadratic forms in standard normal variates. Saddlepoint approximations typically require that the cumulant generating function (cgf) $K_X(s)$ of the random variable X in question be available in a serviceable form. In such cases, the approximations extend straightforwardly to the mean of N such random variables, the reason being that the cgf of a sum of N i.i.d. copies of X is simply $NK_X(s)$. The random variables to be dealt with here do not, however, permit a closed form for their cgf, and thus require custom treatment.

A saddlepoint approximation for the density of

$$R \equiv \frac{\mathbf{X}'\mathbf{A}_1\mathbf{X}}{\mathbf{X}'\mathbf{A}_2\mathbf{X}}, \quad \mathbf{A}_2 \geq 0, \quad \mathbf{X} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_T), \tag{2.23}$$

$\mathbf{A}_1, \mathbf{A}_2$ symmetric, is given by (Lieberman, 1994a)

$$\hat{f}_1(r) = \frac{\phi(\hat{w}) \operatorname{tr} \mathbf{K}_2}{\sqrt{2 \operatorname{tr} \mathbf{K}_3^2}}, \tag{2.24}$$

where $\hat{w} = \sqrt{\log|\mathbf{D}|}\,\mathrm{sgn}(r-\mu)$, $\mathbf{D} = \mathbf{I} - 2\hat{s}\mathbf{A}_3$, $\mathbf{A}_3 = \mathbf{A}_1 - r\mathbf{A}_2$, $\mu = \mathrm{tr}\,\mathbf{A}_1/\mathrm{tr}\,\mathbf{A}_2$, $\mathbf{K}_i = \mathbf{A}_i\mathbf{D}^{-1}$, $i \in \{2,3\}$, ϕ denotes the standard normal pdf, and, with λ_i denoting the eigenvalues of $\mathbf{A}_3$, the saddlepoint $\hat{s}$ solves

$$\mathrm{tr}\,\mathbf{K}_3 = \sum_{i=1}^{T} \frac{\lambda_i}{1-2\hat{s}\lambda_i} = 0. \tag{2.25}$$

Interest centers on the distribution of the mean $\bar{R}$ of N identical and independent copies of R. For ease of exposition, we first consider the random variable

$$S = \sum_{i=1}^{N} R_i,$$

where each of the R_i is identically and independently distributed as (2.23).

Tierney et al. (1989a) show that for an N-dimensional random vector Y having density proportional to

$$f(\mathbf{y}) = b(\mathbf{y})\exp\left(-H(\mathbf{y})\right),$$

a saddlepoint approximation to the marginal density of a (sufficiently smooth) scalar function $x = g(\mathbf{y})$ is given by

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}}\left[\frac{|H''(\hat{\mathbf{y}})|}{|H''(\hat{\mathbf{y}}_x)|\,\nabla g(\hat{\mathbf{y}}_x)' H''(\hat{\mathbf{y}}_x)^{-1}\nabla g(\hat{\mathbf{y}}_x)}\right]^{1/2}\frac{f(\hat{\mathbf{y}}_x)}{f(\hat{\mathbf{y}})}, \tag{2.26}$$

where H'' denotes the Hessian of H, $\hat{\mathbf{y}}$ is a local minimizer of H, and $\hat{\mathbf{y}}_x$ minimizes H subject to $g(\mathbf{y}) = x$. In the i.i.d. case, the joint density takes the form

$$f(\mathbf{y}) = \prod_i f_1(y_i) = \left[\prod_i b_1(y_i)\right]\exp\left(-\sum_i h(y_i)\right),$$

and the Hessian becomes

$$H''(\mathbf{y}) = \mathrm{diag}(h''(y_1),\ldots,h''(y_N)).$$

If $g(\mathbf{y}) = \sum_i y_i$, then $\nabla g(\mathbf{y}) = (1,\ldots,1)$, and, under certain regularity conditions, $\hat{\mathbf{y}}_x = (x/N,\ldots,x/N)$. Furthermore, $\hat{\mathbf{y}} = (\hat{y}_1,\ldots,\hat{y}_1)$, and we can write

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}}\left[\frac{h''(\hat{y}_1)^N}{h''(x/N)^{N-1}N}\right]^{1/2}\frac{f_1(x/N)^N}{f_1(\hat{y}_1)^N}.$$

Taking $f_1(y_i) \equiv \hat{f}_1(y_i)$, i.e., the saddlepoint density given in (2.24), we have $h(y_i) = \frac{1}{2}\log|\mathbf{D}|$. Note that matrix $\mathbf{D}$, as well as $\mathbf{A}_3$, $\mathbf{K}_i$, and quantities $\hat{s}$ and $\hat{w}$, depend on the argument of $\hat{f}_1$; however, in a slight abuse of notation, we shall not make this explicit. We then have

$$h'(y_i) = \frac{1}{2}\,\mathrm{tr}\,\mathbf{D}^{-1}\left(-2\frac{\partial\hat{s}}{\partial y_i}\mathbf{A}_3 + 2\hat{s}\mathbf{A}_2\right) = \hat{s}\,\mathrm{tr}\,\mathbf{K}_2 \tag{2.27}$$

and

$$\begin{aligned} h''(y_i) &= \frac{\partial \hat{s}}{\partial y_i} \operatorname{tr} \mathbf{K}_2 - \hat{s} \operatorname{tr} \mathbf{A}_2 \mathbf{D}^{-1} (-2 \frac{\partial \hat{s}}{\partial y_i} \mathbf{A}_3 + 2\hat{s} \mathbf{A}_2) \mathbf{D}^{-1} \\ &= \frac{\partial \hat{s}}{\partial y_i} \operatorname{tr} \mathbf{K}_2 + 2\hat{s} \frac{\partial \hat{s}}{\partial y_i} \operatorname{tr} \mathbf{K}_2 \mathbf{K}_3 - 2\hat{s}^2 \operatorname{tr} \mathbf{K}_2^2. \end{aligned}$$

Differentiating (2.25),

$$\frac{\partial \hat{s}}{\partial y_i} = \left(2\hat{s} \operatorname{tr} \mathbf{K}_2 \mathbf{K}_3 + \operatorname{tr} \mathbf{K}_2\right) / (2 \operatorname{tr} \mathbf{K}_3^2), \tag{2.28}$$

and, plugging in,

$$h''(y_i) = 2 \left(\frac{\partial \hat{s}}{\partial y_i} \right)^2 \operatorname{tr} \mathbf{K}_3^2 - 2\hat{s}^2 \operatorname{tr} \mathbf{K}_2^2.$$

For the case at hand, $\hat{y}_1 = \mu$, where $\hat{s} = 0$, $\hat{w} = 0$, $\mathbf{K}_2 = \mathbf{A}_2$, and $\mathbf{K}_3 = \mathbf{A}_1 - \mu \mathbf{A}_2$, so that

$$\hat{f}_1(\hat{y}_1) = \frac{1}{\sqrt{2\pi}} \frac{\operatorname{tr} \mathbf{A}_2}{\sqrt{2 \operatorname{tr} \left((\mathbf{A}_1 - \mu \mathbf{A}_2)^2 \right)}}.$$

Also,

$$h''(\hat{y}_1) = \left(\operatorname{tr} \mathbf{A}_2 \right)^2 / \left(2 \operatorname{tr} \left((\mathbf{A}_1 - \mu \mathbf{A}_2)^2 \right) \right),$$

and we obtain

$$\hat{f}(s) = \left[\frac{(2\pi)^{N-1}}{h''(s/N)^{N-1} N} \right]^{1/2} \hat{f}_1(s/N)^N$$

for the sum of ratios. Similarly, for the mean $\bar{R} = S/N$, a univariate transformation shows that

$$\begin{aligned} \hat{f}(\bar{r}) &= \left[\frac{N(2\pi)^{N-1}}{h''(\bar{r})^{N-1}} \right]^{1/2} \hat{f}_1(\bar{r})^N \\ &= \left[\frac{N(2\pi)^{N-1}}{h''(\bar{r})^{N-1}} \right]^{1/2} \frac{\phi(\hat{w})^N \operatorname{tr}^N \mathbf{K}_2}{(2 \operatorname{tr} \mathbf{K}_3^2)^{N/2}} \\ &= \sqrt{\frac{N}{2\pi}} \left[\frac{\operatorname{tr} \mathbf{K}_2}{\sqrt{2 \operatorname{tr} \mathbf{K}_3^2}} \right] \left[\frac{\operatorname{tr}^2 \mathbf{K}_2}{\left(2\hat{s} \operatorname{tr} \mathbf{K}_2 \mathbf{K}_3 + \operatorname{tr} \mathbf{K}_2\right)^2 - 4\hat{s}^2 \operatorname{tr} \mathbf{K}_2^2 \operatorname{tr} \mathbf{K}_3^2} \right]^{(N-1)/2} e^{-N\hat{w}^2/2}. \end{aligned}$$

The approximate cdf can be written

$$\begin{aligned} \hat{F}(x) &= \int_{-\infty}^{x} \sqrt{\frac{N}{2\pi}} \left[\frac{\operatorname{tr} \mathbf{K}_2}{\sqrt{2 \operatorname{tr} \mathbf{K}_3^2}} \right] \left[\frac{\operatorname{tr}^2 \mathbf{K}_2}{\left(2\hat{s} \operatorname{tr} \mathbf{K}_2 \mathbf{K}_3 + \operatorname{tr} \mathbf{K}_2\right)^2 - 4\hat{s}^2 \operatorname{tr} \mathbf{K}_2^2 \operatorname{tr} \mathbf{K}_3^2} \right]^{(N-1)/2} e^{-N\hat{w}^2/2} \mathrm{d}\bar{r} \\ &= \sqrt{\frac{N}{2\pi}} \int_{-\infty}^{\hat{w}(x)} q(\hat{w}) e^{-N\hat{w}^2/2} \mathrm{d}\hat{w}, \end{aligned} \tag{2.29}$$

where

$$q(\hat{w}) = \left[\frac{\operatorname{tr}\mathbf{K}_2}{\sqrt{2\operatorname{tr}\mathbf{K}_3^2}}\right]\left[\frac{\operatorname{tr}^2\mathbf{K}_2}{\left(2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3+\operatorname{tr}\mathbf{K}_2\right)^2-4\hat{s}^2\operatorname{tr}\mathbf{K}_2^2\operatorname{tr}\mathbf{K}_3^2}\right]^{(N-1)/2}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\hat{w}}.$$

Temme (1982) shows that the value of an integral of the form (2.29) is approximately

$$\hat{F}(x) \approx q(0)\Phi\Big(\sqrt{N}\hat{w}(x)\Big) - \frac{q(\hat{w}(x))-q(0)}{\sqrt{N}\hat{w}(x)}\phi\Big(\sqrt{N}\hat{w}(x)\Big),$$

where Φ is the standard normal cdf. It is easily seen that $\frac{\mathrm{d}\hat{w}}{\mathrm{d}\bar{r}} = \frac{\hat{s}\operatorname{tr}\mathbf{K}_2}{\hat{w}}$, so that

$$q(\hat{w}(x)) = \left[\frac{\hat{w}(x)}{\hat{s}\sqrt{2\operatorname{tr}\mathbf{K}_3^2}}\right]\left[\frac{\operatorname{tr}^2\mathbf{K}_2}{\left(2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3+\operatorname{tr}\mathbf{K}_2\right)^2-4\hat{s}^2\operatorname{tr}\mathbf{K}_2^2\operatorname{tr}\mathbf{K}_3^2}\right]^{(N-1)/2},$$

and by applying l'Hôpital's rule twice to the quantity $\frac{\hat{s}\operatorname{tr}\mathbf{K}_2}{\hat{w}}$, we find $q(0) = 1$. By defining

$$\hat{w}_n = \sqrt{N\log|\mathbf{D}|}\,\operatorname{sgn}(\bar{r}-\mu)$$

and

$$\hat{u}_n \equiv \frac{\hat{w}_n}{q(\hat{w})} = \hat{s}\sqrt{2N\operatorname{tr}\mathbf{K}_3^2}\left[\frac{\left(2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3+\operatorname{tr}\mathbf{K}_2\right)^2-4\hat{s}^2\operatorname{tr}\mathbf{K}_2^2\operatorname{tr}\mathbf{K}_3^2}{\operatorname{tr}^2\mathbf{K}_2}\right]^{(N-1)/2},$$

and writing $\bar{r}$ instead of x in the final result, the approximation can be expressed in the more familiar form

$$\hat{F}(\bar{r}) = \Phi(\hat{w}_n) + \phi(\hat{w}_n)\Big(\frac{1}{\hat{w}_n}-\frac{1}{\hat{u}_n}\Big).$$

However, for the case at hand, an alternative form of the approximate cdf turns out to be more reliable: it is given by

$$\hat{F}^*(\bar{r}) = \Phi\left(w_n^*\right), \quad w_n^* \equiv \hat{w}_n + \frac{1}{\hat{w}_n}\log\frac{\hat{u}_n}{\hat{w}_n}, \tag{2.30}$$

and is the approximation used throughout the paper. This form of the approximation is originally due to Barndorff-Nielsen (1986, 1990); its use in the present context can be justified by Lemma 2.1 of Jensen (1992). Note that, at $\bar{r} = \mu$, $\hat{u}_n = \hat{w}_n = 0$, so that (2.30) is not meaningful, and w_n^* should be replaced by its limit, which is given by

$$\lim_{\bar{r}\to\mu} w_n^* = \sqrt{\frac{2}{N\operatorname{tr}\mathbf{A}_{3,\mu}^2}}\left[(N-1)\frac{\operatorname{tr}\mathbf{A}_2\mathbf{A}_{3,\mu}}{\operatorname{tr}\mathbf{A}_2}+\frac{\operatorname{tr}\mathbf{A}_{3,\mu}^3}{3\operatorname{tr}\mathbf{A}_{3,\mu}^2}\right]. \tag{2.31}$$

This can be shown as follows. Since $\frac{\hat{w}_n}{\hat{u}_n} = q(\hat{w})$ and $q\left(\hat{w}(\mu)\right) = q(0) = 1$,

$$\lim_{\bar{r}\to\mu}\left(\frac{1}{\hat{w}}\log\frac{\hat{u}_n}{\hat{w}_n}\right) = \lim_{\bar{r}\to\mu}\frac{-\log q(\hat{w}_n)}{\hat{w}_n}\left(=\frac{0}{0}\right).$$

Using l'Hôpital's rule,

$$\begin{aligned}\lim_{\bar{r}\to\mu}\left(\frac{1}{\hat{w}_n}\log\frac{\hat{u}_n}{\hat{w}_n}\right) &= \lim_{\bar{r}\to\mu}\frac{\hat{u}_n'/\hat{u}_n-\hat{w}_n'/\hat{w}_n}{\hat{w}_n'} = \lim_{\bar{r}\to\mu}\frac{\hat{u}_n'\hat{w}_n-\hat{u}_n\hat{w}_n'}{\hat{u}_n\hat{w}_n\hat{w}_n'}\left(=\frac{0}{0}\right)\\ &\overset{l'H}{=}\lim_{\bar{r}\to\mu}\frac{\hat{u}_n''\hat{w}_n-\hat{u}_n\hat{w}_n''}{\hat{u}_n'\hat{w}_n\hat{w}_n'+\hat{u}_n(\hat{w}_n')^2+\hat{u}_n\hat{w}_n\hat{w}_n''}\left(=\frac{0}{0}\right)\\ &\overset{l'H}{=}\lim_{\bar{r}\to\mu}\frac{\hat{u}_n'''\hat{w}_n+\hat{u}_n''\hat{w}_n'-\hat{u}_n'\hat{w}_n''-\hat{u}_n\hat{w}_n'''}{\hat{u}_n\hat{w}_n\hat{w}_n'''+3\hat{u}_n\hat{w}_n'\hat{w}_n''+2\hat{u}_n'\hat{w}_n\hat{w}_n''+2\hat{u}_n'(\hat{w}_n')^2+\hat{u}_n''\hat{w}_n\hat{w}_n'}\\ &=\lim_{\bar{r}\to\mu}\frac{\hat{u}_n''\hat{w}_n'-\hat{u}_n'\hat{w}_n''}{2\hat{u}_n'(\hat{w}_n')^2},\end{aligned}$$

or, as $1=\lim_{\bar{r}\to\mu}\frac{\hat{u}_n}{\hat{w}_n}\overset{l'H}{=}\lim_{\bar{r}\to\mu}\frac{\hat{u}_n'}{\hat{w}_n'}$,

$$\lim_{\bar{r}\to\mu}\left(\frac{1}{\hat{w}_n}\log\frac{\hat{u}_n}{\hat{w}_n}\right)=\lim_{\bar{r}\to\mu}\frac{\hat{u}_n''-\hat{w}_n''}{2(\hat{u}_n')^2},$$

evaluating which requires an expression for $\lim_{\bar{r}\to\mu}\hat{u}_n'$, $\lim_{\bar{r}\to\mu}\hat{u}_n''$ and $\lim_{\bar{r}\to\mu}\hat{w}_n''$. Straightforwardly,

$$\hat{u}_n'=\hat{s}'\,(*)+\hat{s}\,(*)',$$

where

$$(*)=\sqrt{2N\operatorname{tr}\mathbf{K}_3^2}\left[\frac{(2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3+\operatorname{tr}\mathbf{K}_2)^2-4\hat{s}^2\operatorname{tr}\mathbf{K}_2^2\operatorname{tr}\mathbf{K}_3^2}{\operatorname{tr}^2\mathbf{K}_2}\right]^{(N-1)/2}.$$

It easy to see that $\lim_{\bar{r}\to\mu}\mathbf{K}_3=\mathbf{A}_{3,\mu}$, where $\mathbf{A}_{3,\mu}=\mathbf{A}_1-\mu\mathbf{A}_2$. Thus,

$$\lim_{\bar{r}\to\mu}\hat{u}_n'=\lim_{\bar{r}\to\mu}(\hat{s}')\sqrt{2N\operatorname{tr}\mathbf{A}_{3,\mu}^2}\,.$$

From (2.28) and using $\lim_{\bar{r}\to\mu}\mathbf{K}_2=\mathbf{A}_2$,

$$\begin{aligned}\lim_{\bar{r}\to\mu}\hat{s}' &= \lim_{\bar{r}\to\mu}\frac{2\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3+\operatorname{tr}\mathbf{K}_2}{2\operatorname{tr}\mathbf{K}_3^2}\\ &=\frac{\operatorname{tr}\mathbf{A}_2}{2\operatorname{tr}\mathbf{A}_{3,\mu}^2},\end{aligned}$$

so that

$$\lim_{\bar{r}\to\mu}\hat{u}_n'=\operatorname{tr}\mathbf{A}_2\sqrt{\frac{N}{2\operatorname{tr}\mathbf{A}_{3,\mu}^2}}\,.$$

It easily follows from (2.27) that

$$\hat{w}_n'=\frac{N\hat{s}\operatorname{tr}\mathbf{K}_2}{\hat{w}_n},$$

and thus

$$\hat{w}_n''=\frac{N\hat{s}'\operatorname{tr}\mathbf{K}_2\hat{w}_n+N\hat{s}\operatorname{tr}\dot{\mathbf{K}}_2\hat{w}_n-N\hat{s}\operatorname{tr}\mathbf{K}_2\hat{w}_n'}{\hat{w}_n^2},$$

where a matrix with a dot denotes the elementwise derivative. In the limit,

$$\begin{aligned}
\lim_{\bar{r}\to\mu} \hat{w}_n'' &= \lim_{\bar{r}\to\mu} \frac{N\hat{s}' \operatorname{tr}\mathbf{K}_2\hat{w}_n + N\hat{s}\operatorname{tr}\dot{\mathbf{K}}_2\hat{w}_n - N\hat{s}\operatorname{tr}\mathbf{K}_2\hat{w}_n'}{\hat{w}_n^2} \left(= \frac{0}{0}\right) \\
&= \lim_{\bar{r}\to\mu} \frac{N\hat{s}' \operatorname{tr}\mathbf{K}_2 + N\hat{s}\operatorname{tr}\dot{\mathbf{K}}_2 - (\hat{w}_n')^2}{\hat{w}_n} \\
&\overset{l'H}{=} \lim_{\bar{r}\to\mu} \frac{N\hat{s}'' \operatorname{tr}\mathbf{K}_2 + 2N\hat{s}' \operatorname{tr}\dot{\mathbf{K}}_2 + N\hat{s}\operatorname{tr}\ddot{\mathbf{K}}_2 - 2\hat{w}_n'\hat{w}_n''}{\hat{w}_n'} \\
&= \lim_{\bar{r}\to\mu} \frac{N\hat{s}'' \operatorname{tr}\mathbf{K}_2 + 2N\hat{s}' \operatorname{tr}\dot{\mathbf{K}}_2}{\hat{w}_n'} - 2\lim_{\bar{r}\to\mu} \hat{w}_n'' \\
\Rightarrow \lim_{\bar{r}\to\mu} \hat{w}_n'' &= \lim_{\bar{r}\to\mu} \frac{N\hat{s}'' \operatorname{tr}\mathbf{K}_2 + 2N\hat{s}' \operatorname{tr}\dot{\mathbf{K}}_2}{3\hat{w}_n'}.
\end{aligned}$$

where $\dot{\mathbf{K}}_2 = 2\hat{s}'\mathbf{K}_2\mathbf{K}_3 - 2\hat{s}\mathbf{K}_2$, and therefore $\lim_{\bar{r}\to\mu}\dot{\mathbf{K}}_2 = \frac{\operatorname{tr}\mathbf{A}_2}{\operatorname{tr}\mathbf{A}_{3,\mu}^2}\mathbf{A}_2\mathbf{A}_{3,\mu}$.

Differentiating (2.28),

$$\begin{aligned}
\hat{s}'' =& \frac{4\hat{s}' \operatorname{tr}\mathbf{K}_2\mathbf{K}_3 \operatorname{tr}\mathbf{K}_3^2 + 4\hat{s}(\operatorname{tr}\dot{\mathbf{K}}_2\mathbf{K}_3 + \operatorname{tr}\mathbf{K}_2\dot{\mathbf{K}}_3^2)\operatorname{tr}\mathbf{K}_3^2 + 2\operatorname{tr}\dot{\mathbf{K}}_2 \operatorname{tr}\mathbf{K}_3^2}{4(\operatorname{tr}\mathbf{K}_3^2)^2} \\
&- \frac{8\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3 \operatorname{tr}\mathbf{K}_3\dot{\mathbf{K}}_3 + 4\operatorname{tr}\mathbf{K}_2 \operatorname{tr}\mathbf{K}_3\dot{\mathbf{K}}_3}{(4\operatorname{tr}\mathbf{K}_3^2)^2},
\end{aligned}$$

and, using $\mathbf{K}_3\dot{\mathbf{K}}_3 = -\mathbf{K}_3\mathbf{K}_2 + 2\hat{s}'\mathbf{K}_3^3 - 2\hat{s}\mathbf{K}_3^2\mathbf{K}_2$,

$$\begin{aligned}
\hat{s}'' =& \frac{4\hat{s}' \operatorname{tr}\mathbf{K}_2\mathbf{K}_3 \operatorname{tr}\mathbf{K}_3^2 + 4\hat{s}(\operatorname{tr}\dot{\mathbf{K}}_2\mathbf{K}_3 + \operatorname{tr}\mathbf{K}_2\dot{\mathbf{K}}_3^2)\operatorname{tr}\mathbf{K}_3^2 + 4\hat{s}' \operatorname{tr}\mathbf{K}_2\mathbf{K}_3 \operatorname{tr}\mathbf{K}_3^2 - 4\hat{s}\operatorname{tr}\mathbf{K}_2 \operatorname{tr}\mathbf{K}_3^2}{4(\operatorname{tr}\mathbf{K}_3^2)^2} \\
&- \frac{8\hat{s}\operatorname{tr}\mathbf{K}_2\mathbf{K}_3 \operatorname{tr}\mathbf{K}_3\dot{\mathbf{K}}_3 - 4\operatorname{tr}\mathbf{K}_2 \operatorname{tr}\mathbf{K}_3\mathbf{K}_2 + 8\hat{s}' \operatorname{tr}\mathbf{K}_2 \operatorname{tr}\mathbf{K}_3^3 - 8\hat{s}\operatorname{tr}\mathbf{K}_2 \operatorname{tr}\mathbf{K}_3^2\mathbf{K}_2}{4(\operatorname{tr}\mathbf{K}_3^2)^2}.
\end{aligned}$$

In the limit, using the limiting expressions for $\mathbf{K}_2$, $\mathbf{K}_3$, and $\hat{s}'$,

$$\begin{aligned}
\lim_{\bar{r}\to\mu} \hat{s}'' =& \frac{2\lim_{\bar{r}\to\mu}(\hat{s}')\operatorname{tr}\mathbf{A}_2\mathbf{A}_{3,\mu}\operatorname{tr}\mathbf{A}_{3,\mu}^2 + \operatorname{tr}\mathbf{A}_2\operatorname{tr}\mathbf{A}_{3,\mu}\mathbf{A}_2 - 2\lim_{\bar{r}\to\mu}(\hat{s}')\operatorname{tr}\mathbf{A}_2\operatorname{tr}\mathbf{A}_{3,\mu}^3}{(\operatorname{tr}\mathbf{A}_{3,\mu}^2)^2} \\
=& \frac{2\operatorname{tr}\mathbf{A}_2\operatorname{tr}\mathbf{A}_{3,\mu}\mathbf{A}_2}{(\operatorname{tr}\mathbf{A}_{3,\mu}^2)^2} - \frac{(\operatorname{tr}\mathbf{A}_2)^2\operatorname{tr}\mathbf{A}_{3,\mu}^3}{(\operatorname{tr}\mathbf{A}_{3,\mu}^2)^3}.
\end{aligned}$$

We now require an expression for $\lim_{\bar{r}\to\mu}\hat{u}_n''$. It is immediate that

$$\begin{aligned}
\hat{u}_n'' &= \hat{s}''(*) + 2\hat{s}'(*)' + \hat{s}(*)'' \\
\Rightarrow \lim_{\bar{r}\to\mu}\hat{u}_n'' &= \lim_{\bar{r}\to\mu}(\hat{s}'')\sqrt{2N\operatorname{tr}\mathbf{A}_{3,\mu}^2} + \frac{\operatorname{tr}\mathbf{A}_2}{\operatorname{tr}\mathbf{A}_{3,\mu}^2}\lim_{\bar{r}\to\mu}(*)',
\end{aligned}$$

and straightforward but tedious calculations reveal that

$$\begin{aligned}\lim_{\bar{r}\to\mu}(*)' &= \sqrt{2N}\lim_{\bar{r}\to\mu}\left[\frac{\operatorname{tr}\mathbf{K}_3\dot{\mathbf{K}}_3}{\sqrt{\operatorname{tr}\mathbf{K}_3^2}} + (N-1)\sqrt{\operatorname{tr}\mathbf{K}_3^2}\frac{\operatorname{tr}\dot{\mathbf{K}}_2}{\operatorname{tr}\mathbf{K}_2}\right] \\ &= \sqrt{\frac{2N}{\operatorname{tr}\mathbf{A}_{3,\mu}^2}}\left[\frac{\operatorname{tr}\mathbf{A}_2\operatorname{tr}\mathbf{A}_{3,\mu}^3}{\operatorname{tr}\mathbf{A}_{3,\mu}^2} + (N-2)\operatorname{tr}\mathbf{A}_2\mathbf{A}_{3,\mu}\right].\end{aligned}$$

Plugging in and simplifying yields (2.31).

An important special case occurs if for some conformable matrix $\mathbf{G}$,

$$\mathbf{A}_2\mathbf{A}_2 = \mathbf{A}_2 \quad \text{and } \mathbf{A}_1 = \mathbf{A}_2\mathbf{G}\mathbf{A}_2. \tag{2.32}$$

This constellation appears not only in our ML-based quantile unbiased estimation procedure, but also in the computation of, e.g., the null distribution of the Durbin-Watson statistic. In this case, we have that $\operatorname{tr}\mathbf{K}_2 = \operatorname{rank}\mathbf{A}_2$ and $\operatorname{tr}\mathbf{K}_2\mathbf{K}_3 = 2\hat{s}\operatorname{tr}\mathbf{K}_3^2$, so that $\hat{u}_n$ and $\lim_{\bar{r}\to\mu} w_n^*$ simplify to

$$\hat{u}_n = \hat{s}\sqrt{2N\operatorname{tr}\mathbf{K}_3^2}\left[\frac{4\hat{s}^2\operatorname{tr}\mathbf{K}_3^2 + \operatorname{rank}\mathbf{A}_2}{\operatorname{rank}\mathbf{A}_2}\right]^{(N-1)/2}$$

and

$$\lim_{\bar{r}\to\mu} w_n^* = \sqrt{\frac{2}{N}}\frac{\operatorname{tr}\mathbf{A}_{3,\mu}^3}{3\left[\operatorname{tr}\mathbf{A}_{3,\mu}^2\right]^{3/2}},$$

respectively. It is further computationally advantageous to note that the k nonzero eigenvalues λ_i of $\mathbf{A}_3$ in (2.25) satisfy

$$\lambda_i = \omega_i - \bar{r},$$

where ω_i are the nonzero eigenvalues of $\mathbf{A}_1$. In terms of the λ_i,

$$\hat{w}_n = \operatorname{sgn}(\hat{s})\sqrt{-N\sum_{i=1}^{T}\log\hat{\nu}_i}, \quad \hat{u}_n = \hat{s}\sqrt{2N\sum_{i=1}^{T}\lambda_i^2\hat{\nu}_i^2\left[1 + 4\hat{s}^2k^{-1}\sum_{i=1}^{T}\lambda_i^2\hat{\nu}_i^2\right]^{N-1}},$$

where $\hat{\nu}_i = (1 - 2\hat{s}\lambda_i)^{-1}$, $i \in \{1, \ldots, T\}$.

The aforementioned case is also the one where the approximation performs most favorably. This can be attributed to the fact that (2.26) is based on a Laplace approximation to the marginalizing integral over the joint density, which assumes that the dominant contribution to the integral is from a neighborhood of the maximum of the integrand, and that the location of this maximum is dominated by the exponential term. Now in our application, the maximum of the exponential occurs at $\mu = \operatorname{tr}\mathbf{A}_1/\operatorname{tr}\mathbf{A}_2$, where $\hat{w} = 0$. If $\mathbf{A}_1$ and $\mathbf{A}_2$ satisfy the above conditions, then μ coincides with the mean of $\bar{R}$, which, to order N^{-1}, maximizes the joint density.

The problems arising in other cases (lower accuracy, and the pdf is poorly normalized) could

potentially be remedied by incorporating the term $\operatorname{tr}\mathbf{K}_2/\sqrt{2\operatorname{tr}\mathbf{K}_3^2}$ into the exponent of the component saddlepoint density, i.e.,

$$\hat{f}_1(r) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\hat{w}^2 - \frac{1}{2}\log\operatorname{tr}\mathbf{K}_3^2 + \log\operatorname{tr}\mathbf{K}_2\right\}, \tag{2.33}$$

and proceeding from there. This is what Tierney et al. (1989b) refer to as a fully exponential Laplace approximation. We shall not pursue this here, as, besides leading to less compact formulae, it would entail having to numerically find the mode of (2.33), where for every evaluation of (2.33), the saddlepoint equation (2.25) would have to be solved, thus eroding the time savings otherwise associated with the use of the approximation.

In order to exemplify the virtues of our proposed approximation, we consider the mean of N Durbin-Watson statistics. This can be seen as a version of Bhargava et al. (1982)'s test for serial correlation in panels, adapted for the case of unequal individual variances. We need to evaluate

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{U}'\mathbf{R}_1'\mathbf{M}_1\mathbf{A}\mathbf{M}_1\mathbf{R}_1\mathbf{U}}{\mathbf{U}'\mathbf{R}_1'\mathbf{M}_1\mathbf{R}_1\mathbf{U}} < \bar{r}\right),$$

where the $T+1\times T+1$ matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix},$$

where $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, and $\mathbf{X}$ is a $T+1\times k_1$ regressor matrix. Under the null of no serial correlation, $\mathbf{R}_1 = \mathbf{I}$, and the statistic is precisely in the form (2.32). Under the alternative of a serial correlation of α, $\mathbf{R}_1$ is as in (2.35).

For a model with $T+1 = 25$ and a constant and a time trend as regressors, the simulated and saddlepoint null distributions of the mean of $N \in \{1, 10, 50\}$ independent copies of the Durbin-Watson statistic depicted in the left panel of Figure 2.4 are graphically almost indistinguishable. The right panel of the same figure shows the distribution under the alternative hypothesis that $\alpha = 0.95$, demonstrating the loss of accuracy incurred when the statistic at hand is not in the form (2.32).

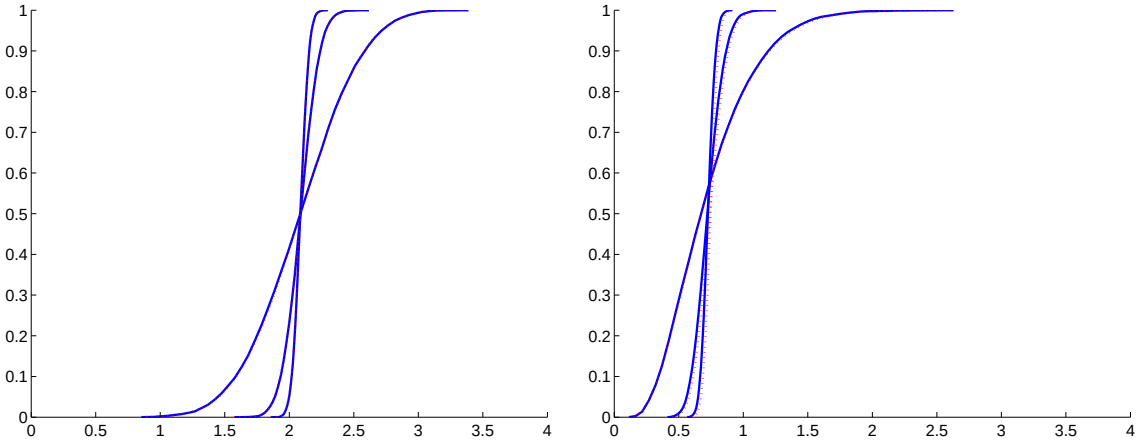

Figure 2.4: Empirical (solid) and saddlepoint (dotted) cdf of the mean of $N \in \{1, 10, 50\}$ Durbin-Watson statistics under $H_0 : \alpha = 0$ (left panel) and $H_1 : \alpha = 0.95$ (right panel). The empirical cdf was obtained by simulation with 10.000 replications.

Appendix 2.B Proofs

$\hat{\alpha}_{\text{LS}}$ *in* (2.10) *has distribution free of* $(\boldsymbol{\beta}', \sigma)$*.*

Neglecting the common factor restrictions on the $2k$ parameters in $\boldsymbol{\gamma}$, computation of (2.10) requires the $NT \times 2k$ matrix $\mathbf{Z}$ to have full column rank, which may fail to hold, e.g., if the regressors include N individual dummies and T time dummies. Thus, if $r = \text{rank}(\mathbf{Z}) < 2k$, replace $\mathbf{Z}$ by

$$\tilde{\mathbf{Z}} = \mathbf{QW},$$

where $\mathbf{QWV}'$ is the singular value decomposition of $\mathbf{Z}$, i.e., $\mathbf{Q}$ and $\mathbf{V}$ are $NT \times r$ and $2k \times r$ matrices, respectively, of full column rank r and $\mathbf{W}$ is an $r \times r$ diagonal matrix of full rank, such that $\mathbf{Z} = \mathbf{QWV}'$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{V}'\mathbf{V} = \mathbf{I}_r$.[2]

Letting $\tilde{\boldsymbol{\gamma}} = \mathbf{V}'\boldsymbol{\gamma}$, rewrite (2.9) as

$$\mathbf{Y} = \alpha\mathbf{Y}_{-1} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\gamma}} + \mathbf{U}, \tag{2.34}$$

and the OLS estimator for α can now be obtained as in (2.10) with $\mathbf{M}$ replaced by $\mathbf{M} = \mathbf{I}_{NT} - \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}' = \mathbf{I}_{NT} - \mathbf{QQ}'$.

To show that $\hat{\alpha}_{\text{LS}}$ does not depend on the value of $\boldsymbol{\beta}$, it suffices to show that the first–step residuals $\mathbf{MY}$ and $\mathbf{MY}_{-1}$ do not depend on $\boldsymbol{\beta}$. Since $\mathbf{MY} = \mathbf{MX}\boldsymbol{\beta} + \mathbf{MY}^{\ell}$ and $\mathbf{MY}_{-1} = \mathbf{MX}_{-1}\boldsymbol{\beta} + \mathbf{MY}^{\ell}_{-1}$, this amounts to showing that $\mathbf{MX} = \mathbf{0}$ and $\mathbf{MX}_{-1} = \mathbf{0}$, because neither $\mathbf{MY}^{\ell}$ nor $\mathbf{MY}^{\ell}_{-1}$ depend on $\boldsymbol{\beta}$. Partitioning $\mathbf{V}'$ into the first k and the last k columns $\mathbf{V}'_1$ and $\mathbf{V}'_2$, respectively, we obtain $\mathbf{Z} = [\mathbf{X}, \mathbf{X}_{-1}] = [\mathbf{QWV}'_1, \mathbf{QWV}'_2]$ and, thus, $\mathbf{X} = \mathbf{QWV}'_1$ and $\mathbf{X}_{-1} = \mathbf{QWV}'_2$. Now, $\mathbf{MX} = (\mathbf{I}_{NT} - \mathbf{QQ}')\mathbf{QWV}'_1 = (\mathbf{Q} - \mathbf{Q})\mathbf{WV}'_1 = \mathbf{0}$ and $\mathbf{MX}_{-1} =$

[2] Clearly, only r different parameters in $\boldsymbol{\gamma}$ are identified.

$(\mathbf{I}_{NT} - \mathbf{QQ}')\mathbf{QWV}_2' = (\mathbf{Q} - \mathbf{Q})\mathbf{WV}_2' = \mathbf{0}$, which proves the invariance of $\hat{\alpha}_{\text{LS}}$ with respect to $\boldsymbol{\beta}$. The invariance of $\hat{\alpha}_{\text{LS}}$ with respect to σ^2 (and to Y_{i0} if $\alpha = 1$) can be proved as outlined by Andrews (1993). ■

Representation of η_1 as a ratio of quadratic forms.

Without loss of generality, we can assume $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$ by virtue of the above invariance property. Defining the $T \times (T+1)$ selection matrices $\mathbf{D}_T = [\mathbf{0} \mid \mathbf{I}_T]$ and $\mathbf{D}_{T-1} = [\mathbf{I}_T \mid \mathbf{0}]$, we have that $\mathbf{MY} = \mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_T\right]\mathbf{Y}_0^\ell$ and $\mathbf{MY}_{-1} = \mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\right]\mathbf{Y}_0^\ell$.

With $\mathbf{U}_0 \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}_{NT+N}\right)$, we have that $\mathbf{Y}_0^\ell = \left[\mathbf{I}_N \otimes \mathbf{R}_1\right]\mathbf{U}_0 \equiv \mathbf{RU}_0$, where

$$\mathbf{R}_1 = \mathbf{R}_1(\alpha) = \begin{bmatrix} b & 0 & 0 & \cdots & 0 & 0 \\ b\alpha & 1 & 0 & \cdots & 0 & 0 \\ b\alpha^2 & \alpha & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b\alpha^T & \alpha^{T-1} & \alpha^{T-2} & \cdots & \alpha & 1 \end{bmatrix}, \tag{2.35}$$

$b = \left(1 - \alpha^2\right)^{-1/2}$ if $\alpha \in (-1, 1)$ and zero if $\alpha = 1$. Substituting this into (2.10) yields

$$\begin{aligned} \eta_1(\alpha) &= \frac{(\mathbf{Y}_0^\ell)'\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\right]'\mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_T\right]\mathbf{Y}_0^\ell}{(\mathbf{Y}_0^\ell)'\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\right]'\mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\right]\mathbf{Y}_0^\ell} - \alpha \\ &= \frac{\mathbf{U}_0'\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\mathbf{R}_1\right]'\mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_T\mathbf{R}_1\right]\mathbf{U}_0}{\mathbf{U}_0'\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\mathbf{R}_1\right]'\mathbf{M}\left[\mathbf{I}_N \otimes \mathbf{D}_{T-1}\mathbf{R}_1\right]\mathbf{U}_0} - \alpha = \frac{\mathbf{U}_0'\mathbf{A}^*\mathbf{U}_0}{\mathbf{U}_0'\mathbf{B}\mathbf{U}_0} - \alpha, \end{aligned}$$

where $\mathbf{A}^*$ and $\mathbf{B}$ are so defined.

Letting $\mathbf{A} = \frac{1}{2}\left[\mathbf{A}^{*\prime} + \mathbf{A}^*\right]$, we can thus write

$$\Pr_\alpha\left(\eta_1(\alpha) \leq \eta_1(\alpha, \boldsymbol{\Upsilon})\right) = \Pr_\alpha\left(\frac{\mathbf{U}_0'\mathbf{A}\mathbf{U}_0}{\mathbf{U}_0'\mathbf{B}\mathbf{U}_0} \leq \frac{\mathbf{Y}_{-1}'\mathbf{MY}}{\mathbf{Y}_{-1}'\mathbf{MY}_{-1}}\right). \tag{2.36}$$

■

Part II

Conditionally Heteroskedastic Models

Manuscript 3

Saddlepoint Approximations for the Doubly Noncentral t Distribution

3.1 Introduction

This paper derives highly accurate and trivially computed approximations of the probability density function (pdf) and cumulative distribution function (cdf) of a doubly noncentral t random variable. Let $T = X/\sqrt{Y/n}$, where X and Y are independent, X has a normal distribution with mean μ and unit variance, and Y has a noncentral χ^2 distribution with $n \in \mathbb{N}$ degrees of freedom, noncentrality parameter θ, and density

$$f_Y(y) = \frac{1}{2} e^{-(y+\theta)/2} y^{(n-2)/4} \theta^{-(n-2)/4} I_{(n-2)/2}\big(\sqrt{\theta y}\big), \quad y > 0,$$

where

$$I_\nu = \sum_{i=0}^{\infty} \frac{(z/2)^{\nu+2i}}{i!\Gamma(\nu + i + 1)}$$

is the modified Bessel function of the first kind of order ν. Random variable T is said to follow a doubly noncentral (Student's) t distribution with n degrees of freedom, numerator noncentrality parameter μ and denominator noncentrality parameter θ. We will write $T \sim t''(n, \mu, \theta)$, with the pdf and cdf of T denoted by $f_{t''}(t; n, \mu, \theta)$ and $F_{t''}(t; n, \mu, \theta)$, respectively. If $\theta = 0$, then T follows a singly noncentral t distribution, denoted $T \sim t'(n, \mu)$. The singly noncentral t is ubiquitous in statistical applications, as it is required for computing the power of a standard t-test. The doubly noncentral t appears quite naturally as the distribution of the t-statistic when the population means are unequal (Robbins, 1948), and in the analysis of variance (Scheffé, 1959, p. 137). Apart from these classical statistical applications, the t-distribution and its noncentral variants have recently been very successfully employed in the modelling of financial returns data; details and references will be given in Section 3.5 below. In such contexts, the domain of the degrees of freedom parameter is typically extended to contain the entire positive half of the real line.

Letting $\omega_{i,\theta} := \exp(-\theta/2)(\theta/2)^i/i!$ and $s_{i,n} := \sqrt{(n+2i)/n}$, Kocherlakota and Kocherlakota (1991) show that

$$f_{t''}(t; n, \mu, \theta) = \sum_{i=0}^{\infty} \omega_{i,\theta} s_{i,n} f_{t'}(s_{i,n} t; n + 2i, \mu) \tag{3.1}$$

and

$$F_{t''}(t; n, \mu, \theta) = \sum_{i=0}^{\infty} \omega_{i,\theta} F_{t'}(s_{i,n} t; n + 2i, \mu), \tag{3.2}$$

where $f_{t'}$ and $F_{t'}$ refer to the singly noncentral t. The former, for example, can be expressed as

$$\begin{aligned} f_{t'}(t; k, \mu) = e^{-\mu^2/2} \frac{\Gamma\left((k+1)/2\right) k^{k/2}}{\sqrt{\pi}\Gamma\left(k/2\right)} \left(\frac{1}{k + t^2}\right)^{\frac{k+1}{2}} \\ \times \left(\sum_{i=0}^{\infty} \frac{(t\mu)^i}{i!} \left(\frac{2}{t^2 + k}\right)^{i/2} \frac{\Gamma\left((k+i+1)/2\right)}{\Gamma\left((k+1)/2\right)} \right). \end{aligned} \tag{3.3}$$

In the doubly noncentral case, computing the doubly infinite sum required for the pdf or cdf of T at a given point t to any specified degree of accuracy is straightforward, but time consuming. As an example, computation of the 401 pdf values plotted in Figure 3.1 required 2.2 seconds using (3.1) on a 2.8 GHz PC, and 0.004 seconds using the approximation developed herein. An example below will demonstrate a situation in which the pdf needs to be evaluated on the order of 10^6 times; use of (3.1) would be practically impossible.

Approximations to the distribution of T were considered in Krishnan (1967) and Mudholkar and Chaubey (1976), which are discussed and compared below, while Krishnan (1968) and Nandi and Choudhury (2002) considered series representations of the distribution function. This paper relies on a result of Daniels and Young (1991) to develop highly accurate saddlepoint approximations, hereafter SPA, to both the density and the distribution function of the doubly (and, hence, singly) noncentral t. We demonstrate the outstanding accuracy of the SPA–based method, which holds over the entire support of T, and compare it to existing approximations, which often exhibit errors which are several orders of magnitude larger than obtained by the new approximation. Indeed, in the central Student's t case, the (renormalized) approximation is exact.

In most cases of practical interest, the SPA requires finding the root of an equation or, possibly, solving a nonlinear system of equations for each ordinate t. This differs from use of, say, an Edgeworth expansion, which yields a closed-form expression amenable to fast "vectorized" computation available in modern matrix–based programming languages such as Splus and Matlab. Interestingly, and quite fortunately, the required system of equations in this case can be solved analytically, yielding a completely closed–form expression and obviating multivariate root searching and the potential numerical problems inherently associated with it. Thus, the pdf and cdf approximations are evaluated essentially instantaneously, and never fail.

The remainder of this paper is organized as follows. Section 3.2 derives the saddlepoint approximation and provides the analytic solution to the requisite system of equations. (That this solution is unique is shown in the appendix.) Section 3.3 provides an improved approximation to the pdf. Section 3.4 contains numerical results on the accuracy of the proposed approximations and offers a comparison with alternative approximations. Finally, Section 3.5 illustrates a useful application of the new method.

3.2 Main Result

Let $\mathbf{x} = (x_1, x_2)$ be a bivariate random vector having a density and a joint cumulant generating function, and denote the latter as $K(\mathbf{t})$. Denote by $K'(\mathbf{t})$ and $K''(\mathbf{t})$ its gradient and hessian, respectively. Consider a bijection $\mathbf{y} = (y_1, y_2) = g^{-1}(\mathbf{x}) = (g_1^{-1}(\mathbf{x}), g_2^{-1}(\mathbf{x}))'$, so that $\mathbf{x} = g(\mathbf{y}) = (g_1(\mathbf{y}), g_2(\mathbf{y}))'$, and let $\nabla_{y_i} g(\mathbf{y}) = (\partial g_1/\partial y_i, \partial g_2/\partial y_i)', i \in \{1, 2\}$. Daniels and Young (1991) show that saddlepoint approximations to the marginal pdf and cdf of y_1 are given by

$$\hat{f}_{Y_1}(y_1) = \phi(w)/u \tag{3.4}$$

and

$$\hat{F}_{Y_1}(y_1) = \Phi(w) + \phi(w)\left(\frac{1}{w} - \frac{d}{u}\right), \tag{3.5}$$

respectively, where ϕ and Φ are the standard normal pdf and cdf, respectively, $w = \sqrt{2\left(\hat{\mathbf{t}}' g(\hat{\mathbf{y}}) - K(\hat{\mathbf{t}})\right)}\,\mathrm{sgn}(y_1 - \alpha)$, $\hat{\mathbf{y}} = (y_1, \hat{y}_2)$, $\alpha = g_1^{-1}(K'(\mathbf{0}))$, $d = \left(\hat{\mathbf{t}}' \nabla_{y_1} g(\hat{\mathbf{y}})\right)^{-1}$,

$$u = \frac{\sqrt{\det\left(K''(\hat{\mathbf{t}})\right)\left[\nabla_{y_2} g(\hat{\mathbf{y}})' \left(K''(\hat{\mathbf{t}})\right)^{-1} \nabla_{y_2} g(\hat{\mathbf{y}}) + \hat{\mathbf{t}}' \nabla^2_{y_2} g(\hat{\mathbf{y}})\right]}}{\det\left(\partial g / \partial \mathbf{y}(\hat{\mathbf{y}})\right)},$$

and, for each value of y_1, $\hat{\mathbf{t}}$ and $\hat{y}_2$ solve the system

$$\begin{aligned} K'(\hat{\mathbf{t}}) &= g(\hat{\mathbf{y}}) \\ \hat{\mathbf{t}}' \nabla_{y_2} g(\hat{\mathbf{y}}) &= 0. \end{aligned}$$

Paralleling the case of univariate saddlepoint approximations, the pdf approximation can be renormalized by numerically integrating it over its support if additional accuracy is desired.

Now let $x_1 \sim \mathrm{N}(\mu, 1)$, independent of $x_2 \sim \chi^2(n, \theta)$, and let $\mathbf{x} = g(\mathbf{y}) = (y_1 y_2, y_2^2 n)$, so that $(y_1, y_2) = g^{-1}(x_1, x_2) = \left(x_1 / \sqrt{x_2/n}, \sqrt{x_2/n}\right)'$ and $y_1 \sim t''(n, \mu, \theta)$. The joint cumulant generating function of (x_1, x_2) is, from independence,

$$K(\mathbf{t}) = K_{x_1}(t_1) + K_{x_2}(t_2) = t_1\mu + \frac{1}{2}t_1^2 - \frac{n}{2}\log(1 - 2t_2) + \frac{t_2\theta}{1 - 2t_2},$$

where K_{x_1} and K_{x_2} are the cumulant generating functions of x_1 and x_2, respectively. The saddlepoint $(\hat{\mathbf{t}}, \hat{y}_2) = (\hat{t}_1, \hat{t}_2, \hat{y}_2)$, $\hat{t}_2 < \frac{1}{2}$, $\hat{y}_2 > 0$, solves the system of equations

$$\begin{aligned} \mu + t_1 &= y_1 y_2 \\ \frac{n}{(1 - 2t_2)} + \frac{\theta}{(1 - 2t_2)^2} &= n y_2^2 \\ t_1 y_1 + 2n t_2 y_2 &= 0. \end{aligned}$$

Straightforward calculation reveals that

$$\hat{t}_1 = -\mu + y_1 \hat{y}_2, \quad \hat{t}_2 = -\frac{y_1 \hat{t}_1}{2n\hat{y}_2}, \tag{3.6}$$

and $\hat{y}_2$ solves the cubic $s(y_2) := a_3 y_2^3 + a_2 y_2^2 + a_1 y_2 + a_0 = 0$, where

$$a_3 = y_1^4 + 2n y_1^2 + n^2, \quad a_2 = -2y_1^3\mu - 2y_1 n\mu, \quad a_1 = y_1^2\mu^2 - n y_1^2 - n^2 - \theta n,$$

and $a_0 = y_1 n\mu$. Upon defining

$$c_2 = \frac{a_2}{a_3}, \quad c_1 = \frac{a_1}{a_3}, \quad c_0 = \frac{a_0}{a_3}, \quad q = \frac{1}{3}c_1 - \frac{1}{9}c_2^2,$$

$$r = \frac{1}{6}(c_1 c_2 - 3c_0) - \frac{1}{27}c_2^3, \quad m = q^3 + r^2, \quad \text{and} \quad s_{1,2} = (r \pm \sqrt{m})^{1/3},$$

the roots of the cubic are given by

$$z_1 = (s_1 + s_2) - \frac{c_2}{3} \qquad \text{and} \qquad z_{2,3} = -\frac{1}{2}(s_1 + s_2) - \frac{c_2}{3} \pm \frac{i\sqrt{3}}{2}(s_1 - s_2).$$

The saddlepoint solution is always z_1, as proved in the appendix. It can also be expressed as

$$\hat{y}_2 = \sqrt{-4q}\cos\Big(\cos^{-1}\big(r/\sqrt{-q^3}\big)/3\Big) - \frac{c_2}{3}, \tag{3.7}$$

thus avoiding complex arithmetic.

With

$$\nabla_{y_1} g(\mathbf{y}) = (y_2, 0)', \quad \nabla_{y_2} g(\mathbf{y}) = (y_1, 2ny_2)',$$

$$\nabla^2_{y_2} g(\mathbf{y}) = (0, 2n)', \quad K''(\mathbf{t}) = \operatorname{diag}\big(1, 2n(1-2t_2)^{-2} + 4\theta(1-2t_2)^{-3}\big),$$

$$\det\big[\partial g/\partial \mathbf{y}\big] = \det\big[\nabla_{y_1} g(\mathbf{y}),\, \nabla_{y_2} g(\mathbf{y})\big] = 2ny_2^2$$

and after some simplification, the quantities entering approximations (3.4) and (3.5) take the simple form

$$d = (\hat{t}_1 \hat{y}_2)^{-1}, \qquad u = \sqrt{(y_1^2 + 2n\hat{t}_2)(2n\nu^2 + 4\theta\nu^3) + 4n^2\hat{y}_2^2}\Big/\Big(2n\hat{y}_2^2\Big),$$

$$w = \sqrt{-\mu\hat{t}_1 - n\log\nu - 2\theta\nu\hat{t}_2}\,\operatorname{sgn}\big(y_1 - \alpha\big), \qquad \alpha = \mu/\sqrt{1 + \theta/n},$$

where $\nu = (1 - 2\hat{t}_2)^{-1}$, and $\hat{\mathbf{t}} = (\hat{t}_1, \hat{t}_2)'$ and $\hat{y}_2$ are given by (3.6) and (3.7), respectively. In the singly noncentral case with $\theta = 0$, these reduce to

$$u = \sqrt{(\mu y_1 \hat{y}_2 + 2n)/(2n)}/\hat{y}_2, \qquad \text{and} \qquad w = \sqrt{-\mu\hat{t}_1 - 2n\log(\hat{y}_2)}\,\operatorname{sgn}(y_1 - \mu),$$

where d, $\hat{t}_1$ and $\hat{t}_2$ are as before, and

$$\hat{y}_2 = \frac{\mu y_1 + \sqrt{4n(y_1^2 + n) + \mu^2 y_1^2}}{2(y_1^2 + n)}.$$

This is the approximation given in DiCiccio and Martin (1991, p. 897, Eq. (18)). In the central case with $\mu = 0$,

$$\hat{y}_2 = \sqrt{n/(y_1^2 + n)}, \quad d = (y_1 \hat{y}_2^2)^{-1}, \quad u = \hat{y}_2^{-1}, \quad w = \sqrt{-2n\log(\hat{y}_2)}\,\operatorname{sgn}(y_1),$$

and the pdf approximation becomes

$$\hat{f}_t(y_1; n) = \frac{1}{\sqrt{2\pi}} \left(\frac{n}{y_1^2 + n} \right)^{\frac{1}{2}(n+1)},$$

which is exact after renormalization.

Note that $\lim_{y_1 \to \alpha} w = \lim_{y_1 \to \alpha} d^{-1} = 0$, so that at $y_1 = \alpha$, expression (3.5) is not well defined. This singularity is, however, removable, and a repeated application of l'Hôpital's rule shows that the limiting value is given by

$$\lim_{y_1 \to \alpha} \hat{F}_{t''}(y_1, n, \mu, \theta) = \frac{1}{2} - \frac{1}{6\sqrt{\pi}} \frac{\mu\big((n+3\theta)(2\mu^2+3n)+6\theta^2\big)}{\big((n+2\theta)(\mu^2+2n)+2\theta^2\big)^{3/2}}. \tag{3.8}$$

While knowledge of this limiting value ensures that the saddlepoint cdf is continuous everywhere, numerical inaccuracies may arise in the immediate vicinity of α. In practice, these are most easily circumvented by replacing (3.5) with

$$\tilde{F}(y_1) = \lim_{y_1 \to \alpha} \hat{F}_{t''}(y_1, n, \mu, \theta) + (y_1 - \alpha) \lim_{y_1 \to \alpha} \hat{F}'_{t''}(y_1, n, \mu, \theta)$$

whenever $|w| < \epsilon$, and where the second limit is given in (3.10) below. The optimal value of ϵ depends on the arithmetic precision of the machine at hand. The left panel of Figure 3.2 demonstrates that in doing so, little accuracy is lost, owing to the approximate linearity of the cdf near α.

3.3 Avoiding Renormalization

As noted before, the saddlepoint approximation to the pdf can be normalized to integrate to unity, resulting in greater accuracy. Also, in the context of maximum likelihood estimation, use of the unnormalized approximation leads to biased estimates. The reason for this is that the constant of integration,

$$k(n, \mu, \theta) = \left[\int_{-\infty}^{\infty} \hat{f}_{t''}(t; n, \mu, \theta)\, \mathrm{d}t \right]^{-1},$$

depends on the parameters of the distribution; e.g., for the doubly noncentral t distribution considered here, our experiments show that the normalizing constant is an increasing function of all the parameters, but is most sensitive to variations in the degrees of freedom n. Thus, if $\hat{n}$ maximizes the normalized likelihood, then evaluating the non-normalized likelihood at a value $n_1 < \hat{n}$ may spuriously result in a higher likelihood value, simply because the densities are not properly normalized. This problem of downward-biased estimates of n is particularly acute when the degrees of freedom parameter is small, which is precisely the case in financial applications, where fitting the singly noncentral t has become commonplace. (Note that the normalizing constant approaches unity as n grows to infinity.)

However, renormalization of the pdf involves numeric integration, and so considerably slows down the otherwise virtually instantaneous computation of the saddlepoint pdf. In an application with a large number of evaluations of the density, such as occurs when conducting maximum likelihood estimation, this factor becomes noticeable. For example, computing the 401 approximate pdf values of Figure 3.1 takes about 0.056 seconds, or 14 times longer, when using Matlab's built-in `quadl` routine for normalization. (The exact factor by which the computational effort is increased depends on whether the pdf is to be evaluated at many points for the same parameter values, or at a single point for varying values of the parameters. This is due to the fact that the normalizing constant will only have to be evaluated once for each parameter constellation.) This motivates the construction of a pdf approximation which does not require renormalization. We consider two ways.

3.3.1 Analytic: Differentiating the CDF

Routledge and Tsao (1997) show that differentiating the Lugannani and Rice (1980) approximation to the distribution function of the mean of a sample of independent and identically distributed (i.i.d.) random variables gives rise to the same asymptotic expansion as the Daniels (1954) approximation to the corresponding density. Keeping only the terms obtained by differentiating the first-order cdf approximation, a density approximation is obtained that formally has the same error as the first-order pdf approximation, but which integrates to one, thus obviating the need to renormalize.

In the present context, differentiating expression (3.5) for the cdf leads to the adjusted pdf approximation

$$\hat{F}'_{Y_1}(y_1) = \phi(w)\left(\frac{1}{u} - \frac{1}{dw^3} - \frac{d'}{u} + \frac{du'}{u^2}\right), \tag{3.9}$$

where w, d, and u are as before,

$$u' \equiv \frac{\partial}{\partial y_1} u \qquad \text{and} \qquad d' \equiv \frac{\partial}{\partial y_1} d.$$

The general expressions for u' and d' are lengthy and not particularly insightful. However, in our setting, they simplify markedly, due to the independence of the random variables constituting the doubly noncentral t distribution. Upon defining

$$c_1 \equiv 2n\nu^2 + 4\theta\nu^3, \quad c_2 \equiv 8n\nu^3 + 24\theta\nu^4, \quad c_3 \equiv \hat{t}_1 + y_1\hat{y}_2, \quad c_4 \equiv 2n\hat{t}_2 + y_1^2 + c_1^{-1}(2n\hat{y}_2)^2$$

$$c_5 \equiv c_3/c_4, \quad \text{and} \quad c_6 \equiv 2nc_1^{-1}c_2c_5\hat{y}_2(2n\hat{t}_2 + y_1^2) + 12n^2c_5\hat{y}_2 + 2c_1y_1,$$

we obtain

$$d' = -d^2\left(\hat{y}_2^2 - c_3c_5\right) \quad \text{and} \quad u' = -uc_5\left(c_6(2c_1c_3)^{-1} + 2\hat{y}_2^{-1}\right).$$

The adjusted pdf approximation, like the saddlepoint cdf, has a removable singularity at the

point $\mu/\sqrt{1+\theta/n}$. The limiting value can be found analytically, using a derivation similar to the one which led to (3.8) for the cdf. Upon defining $\kappa_i = n + i\theta, i \in \{1, 2, 4\}$, $\omega_1 = 6\kappa_2^2 - \kappa_4$, $\omega_2 = 144\kappa_1^2 - 33\kappa_2$, and $\omega_3 = 4\kappa_1^2 - \kappa_2$, it is given by

$$\lim_{y_1 \to \alpha} \hat{F}'_{t''} = \frac{2\mu^6\big[\kappa_2\omega_1 - 10\theta^2\big] + 12\mu^4\big[\omega_1\kappa_1^2 - 7\theta^2\kappa_1 + \theta^3\big] + \mu^2\big[\kappa_1^2\kappa_2\omega_2\big] + 24\big[\kappa_1^4\omega_3\big]}{12\sqrt{n\pi}\kappa_1^{-3/2}(\kappa_2(\mu^2+2n) + 2\theta^2)^{7/2}}. \tag{3.10}$$

3.3.2 Numeric: Response Surface Fitting

The second way of avoiding numerical integration at run-time is to fit a response surface to pre-computed values of the normalizing constant. In view of the aforementioned ubiquity of the singly noncentral t distribution in financial applications, we restrict our attention to that special case. Our experiments showed that for a given value of μ, $k(n, \mu, 0)$ can be parsimoniously and accurately approximated by regressing $k(n, \mu, 0)$ on several fractional powers of n and $\log(n)$. This process is repeated for a fine grid of values for μ, and the resulting matrix of regression parameters is stored. The approximate normalizing constant, say $\hat{k}(n, \mu)$, is then determined at run-time by linearly interpolating between the least-squares predictions of $\hat{k}(n, \mu, 0)$ for the nearest values of μ.

This response surface was fitted for the range $0.5 \leq n \leq 20$ and $|\mu| \in \{0, 0.1, 0.2, \ldots, 10\}$, which appears adequate for applications in empirical finance. The resulting approximation is accurate to within 1% of the true values. A program written in Matlab is available from the authors to compute all entertained approximations (renormalized via integration, cdf differentiation, and, for the singly noncentral case, the renormalized pdf using the response surface approximation for the integration constant).

3.4 Accuracy Assessment

Several approximations for the cdf have been previously proposed; Krishnan (1967) suggests two different approximations to $F_{t''}(t; n, \mu, \theta)$. The first one is based on a scale transformed singly noncentral t distribution with scale parameter and degrees of freedom chosen to equate the first two moments; it is given by

$$\hat{F}^1_{t''}(t; n, \mu, \theta) = F_{t'}(t/c; f, \mu), \tag{3.11}$$

where

$$f = \frac{7}{2}\left(-1 + \sqrt{15 - 7g^2/h}\right)^{-1}, \quad c = \sqrt{h(1-2/f)}, \quad g = m_1/\sqrt{\mu^2/2}, \quad h = m_2/(1+\mu^2),$$

and m_i denotes the i^{th} raw moment of t'', the first three of which are given by

$$m_1 = \mu\left(\frac{n}{2}\right)^{1/2}\frac{\Gamma\big((n-1)/2\big)}{\Gamma\big(n/2\big)}\,{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, -\frac{\theta}{2}\right), \tag{3.12}$$

$$m_2 = (1+\mu^2)\frac{n}{n-2}\,{}_1F_1\left(\frac{1}{2},\frac{n}{2},-\frac{\theta}{2}\right)$$

and

$$m_3 = \mu(\mu^2+3)\left(\frac{n}{2}\right)^{3/2}\frac{\Gamma\big((n-3)/2\big)}{\Gamma\big(n/2\big)}\,{}_1F_1\left(\frac{3}{2},\frac{n}{2},-\frac{\theta}{2}\right), \tag{3.13}$$

(see Krishnan, 1967), where ${}_1F_1$ is the confluent hypergeometric function. (Note that the expression for m_2 given in both Kocherlakota and Kocherlakota (1991) and Johnson et al. (1995, p. 534) contains an error.)

The second approximation is based on the distribution function of the sample correlation coefficient and is given by

$$\hat{F}^2_{t''}(t;n,\mu,\theta) = F_R(r_0,\rho,n+2), \tag{3.14}$$

where $r_0 = t\big(t^2+nk^2\big)^{-1/2}$,

$$k = \sqrt{K\big/\big(1+(n+1)\rho^2/(1-\rho^2)\big)}, \quad \rho = \sqrt{(3K-L)/(nL-(n-1)K)},$$

$L = (n-3)m_3/(nm_1)$, $K = (1-2/n)m_2$, and, with B and ${}_2F_1$ denoting the beta and hypergeometric function, respectively,

$$F_R(r,\rho,n) = \int_{-1}^{r}\frac{(n-2)(1-\rho^2)^{(n-1)/2}(1-x^2)^{(n-4)/2}}{\sqrt{2}(n-1)B(1/2,n-1/2)(1-\rho x)^{n-3/2}}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2},n-\frac{1}{2},\frac{1}{2}(1+\rho x)\right)\mathrm{d}x \tag{3.15}$$

is the cdf of the correlation coefficient of a size n sample from a bivariate normal distribution with correlation ρ.

The hypergeometric function in the integrand in (3.15) presents a potential problem; its presence renders approximation (3.14) about as time–consuming to evaluate as the exact t'' cdf. This can, however, be remedied by replacing it with its highly accurate Laplace approximation given in Butler and Wood (2002). In particular, they show that

$${}_2\hat{F}_1\,(a,b,c;z) = \frac{{}_2\tilde{F}_1\,(a,b,c;z)}{{}_2\tilde{F}_1\,(a,b,c;0)}, \tag{3.16}$$

where

$${}_2\tilde{F}_1(a,b,c;x) = \lambda^{-1/2}\hat{y}^a\,(1-\hat{y})^{c-a}\,(1-x\hat{y})^{-b}$$

and

$$\begin{aligned}
\lambda &= a\,(1-\hat{y})^2 + (c-a)\,\hat{y}^2 - bx^2\hat{y}^2\,(1-\hat{y})^2\,/\,(1-x\hat{y})^2\,,\\
\hat{y} &= \begin{cases} \left[\tau+\sqrt{\tau^2-4ax\,(c-b)}\right]/\,[2x\,(b-c)]\,, & \text{if } x\neq 0,\\ a/c, & \text{if } x=0,\end{cases}\\
\tau &= x\,(b-a)-c.
\end{aligned}$$

In all our experiments, use of (3.16) resulted in sizeable time savings, without sacrificing accuracy relevant for practical applications.

Given the use of ${}_1F_1$ in (3.13) and below in (3.21), we state the further result from Butler and Wood (2002):

$$ {}_1\hat{F}_1\left(a,b;z\right) = \frac{{}_1\tilde{F}_1\left(a,b;z\right)}{{}_1\tilde{F}_1\left(a,b;0\right)} \tag{3.17} $$

where

$$ \begin{aligned} {}_1\tilde{F}_1(a;b;x) &= \lambda^{-1/2}\hat{y}^a\left(1-\hat{y}\right)^{b-a}e^{x\hat{y}}, \\ \lambda &= a\left(1-\hat{y}\right)^2+(b-a)\,\hat{y}^2, \\ \hat{y} &= \begin{cases} \left[(x-b)+\sqrt{(x-b)^2+4ax}\right]/2x, & \text{if } x\neq 0, \\ a/b, & \text{if } x=0. \end{cases} \end{aligned} $$

Two closed–form approximations to the t'' cdf are given in Mudholkar and Chaubey (1976). The first one is based on Patnaik's (1949) approximation to the noncentral chi square distribution and is given by

$$ \hat{F}^3_{t''}(t;n,\mu,\theta) = \Phi(t^*), \tag{3.18} $$

where $t^* = (\mu_0 t-\mu)\Big/\sqrt{1+\sigma_0^2 t^2}$, $\mu_0 = \sqrt{(n+\theta)/n-\sigma_0^2}$ and $\sigma_0^2 = (n+2\theta)/(2n(n+\theta))$. The second approximation is based on an Edgeworth expansion and is given by

$$ \hat{F}^4_{t''}(t;n,\mu,\theta) = \Phi(z)-\phi(z)\left(\frac{\beta_1}{6}(z^2-1)+\frac{\beta_2}{24}(z^3-3z)+\frac{\beta_1^2}{72}(z^5-10z^3+15z)\right), \tag{3.19} $$

where $z=-\mu_1/\sigma_1$, $\mu_1=\mu-t\kappa_1$, $\sigma_1^2=1+t^2\kappa_2$, $\beta_1=-t^3\kappa_3/\sigma^3$, $\beta_2=t^4\kappa_4/\sigma^4$, and

$$ \begin{aligned} \kappa_1 &= \sqrt{1+\theta/n}\left(1-\frac{1}{2(n+\theta)}+\frac{2n-3}{8(n+\theta)^2}+\frac{14n-15}{16(n+\theta)^3}-\frac{60n^2-580n+525}{128(n+\theta)^4}\right), \\ \kappa_2 &= n^{-1}\left(1-\frac{n-1}{2(n+\theta)}-\frac{3(n-1)}{2(n+\theta)^2}+\frac{7n^2-64n+57}{8(n+\theta)^3}\right), \\ \kappa_3 &= (1+\theta/n)^{3/2}\left(\frac{n-1}{(n+\theta)^3}-\frac{3(n^2-10n+9)}{4(n+\theta)^4}\right), \text{ and} \\ \kappa_4 &= \frac{3(n-1)}{n^2(n+\theta)^2}. \end{aligned} $$

To illustrate the merits of the various approximations, we consider the pdf and cdf of $T\sim t''(n,\mu,\theta)$, for the set of arbitrarily chosen parameters $(n,\mu,\theta)=(5,2,5)$, over a range of ± 10 standard deviations around the mean. The exact pdf and cdf values have been computed from (3.1) and (3.2), respectively, using Matlab's built-in routine for evaluating the singly noncentral pdf and cdf. Denoting the summands in equations (3.1) and (3.2) by S_i^1 and S_i^2, respectively, the infinite sums were truncated at $\bar{i}_j = \inf\{i : i>\theta/2 \ \wedge\ |S_i^j|<2.3\times 10^{-16}\}$, as the absolute

values of the summands S_i^j assume their maximum at some $i \leq \lceil \theta/2 \rceil$, after which they decrease monotonically.

Figure 3.1 demonstrates the accuracy of the pdf approximations. Even for the degrees of freedom as low as $n = 5$, the renormalized and adjusted approximations, shown in the left panel, are graphically indistinguishable from the true pdf. The right panel shows that near the mean, the adjusted pdf approximation has relative percentage error, defined as $100 \times (\text{approx} - \text{exact})/\text{exact}$, comparable to the renormalized approximation; this is true in general, and agrees with the findings of Routledge and Tsao (1997) who show that, near the mean, the differentiated Lugannani–Rice formula is a second order approximation. In the limit as $|t| \to \infty$, $\hat{F}'_{t''}(t)/\hat{f}_{t''}(t) = 1$.

The right-hand panel of Figure 3.2 shows the relative percentage error, now defined as $100 \times (\text{approx} - \text{exact})/\min(\text{exact}, 1 - \text{exact})$, of the various cdf approximations. Although the graph refers only to a particular parameter set, it is to a large extent representative of the general picture: Approximations (3.18) and (3.19) tend to perform acceptably near the mean, but break down for larger deviations. At least for the latter approximation, this behavior is expected, as it is derived as an Edgeworth expansion; these are well-known to deteriorate in the tails of the distribution, see, e.g., Daniels (1954, p. 631).

Approximations (3.11) and (3.14), on the other hand, are useful over the entire real line; however, their relative error is generally several orders of magnitude higher than that of the SPA. It should also be noted that for approximations (3.11) and (3.14) to exist, it is required that $n > 2$ and $n > 4$, respectively, which may not be fulfilled in practice, especially in the context of empirical finance, which we consider below. Also, approximation (3.11) is not defined if $\mu = 0$. Finally, for some combinations of n, μ and θ, the value of ρ exceeds one in (3.14), rendering it useless. The SPA, on the other hand, is well-defined over the entire support of the distribution.

To get an idea how the error from the SPA behaves as a function of the parameters n, μ and θ, Figure 3.3 plots, for several values of θ and as a function of n, the relative percentage error of the saddlepoint approximation to the cdf evaluated at the ordinate t such that $F_{t''}(t; n, \mu, \theta) = p$, where $p = 0.95$ was used. The quantiles for each constellation of points shown were computed by root search using the SPA to the cdf. From the left panel, for which $\mu = 0$, we see that, beyond four degrees of freedom, the approximation has less than one percent error. The right panel is the same, but uses $\mu = 10$, which performs very similarly except for the—not overly practical—case of less than one degree of freedom. A variety of similar runs using different p and μ confirm that Figure 3.3 is quite typical: for larger than four degrees of freedom, the relative error is under one percent, while as n approaches zero, the error increases without bound. Not apparent from the graphs is that, as θ increases past six, the accuracy increases (for example, with $p = 0.95$, $\mu = 0$ and $n = 1$ degree of freedom, the error is already under one percent for $\theta = 12$, and continues to decrease as θ increases).

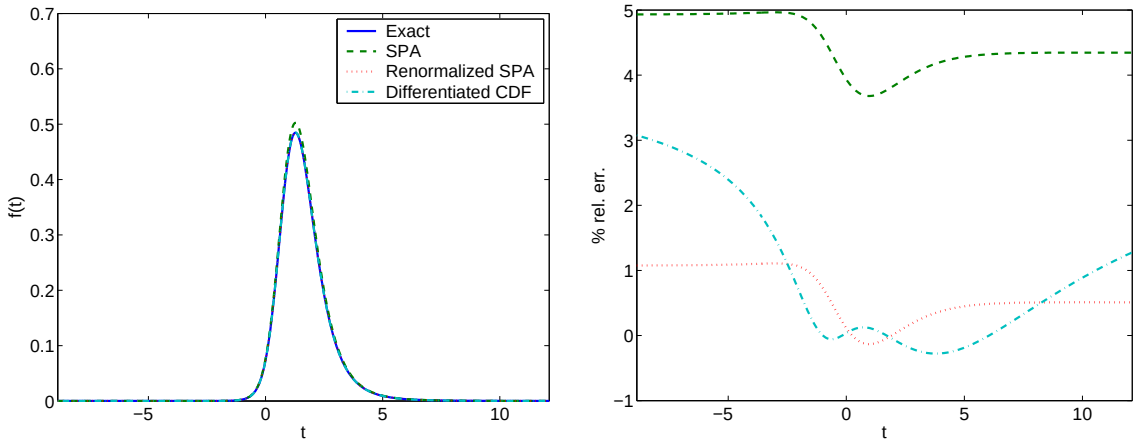

Figure 3.1: Left panel: Density of $T \sim t''(5, 2, 5)$. **Right panel**: Relative percentage errors.

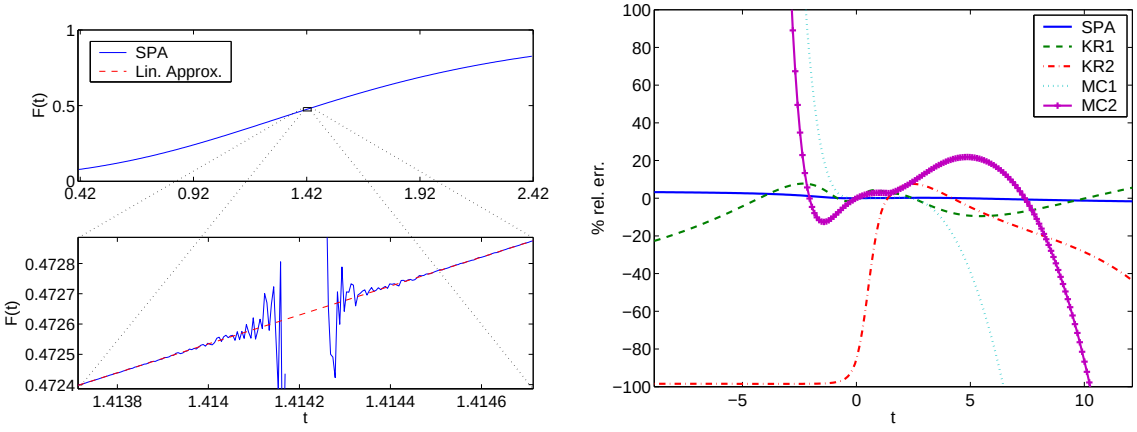

Figure 3.2: Left Panel: Illustration of numerical inaccuracies near α, $T \sim t''(5, 2, 5)$. **Right Panel**: Percentage relative errors of approximations (3.5) (solid), (3.11) (dashes), (3.14) (dash-dot), (3.18) (dots), and (3.19) (crosses), to the cdf of $T \sim t''(5, 2, 5)$.

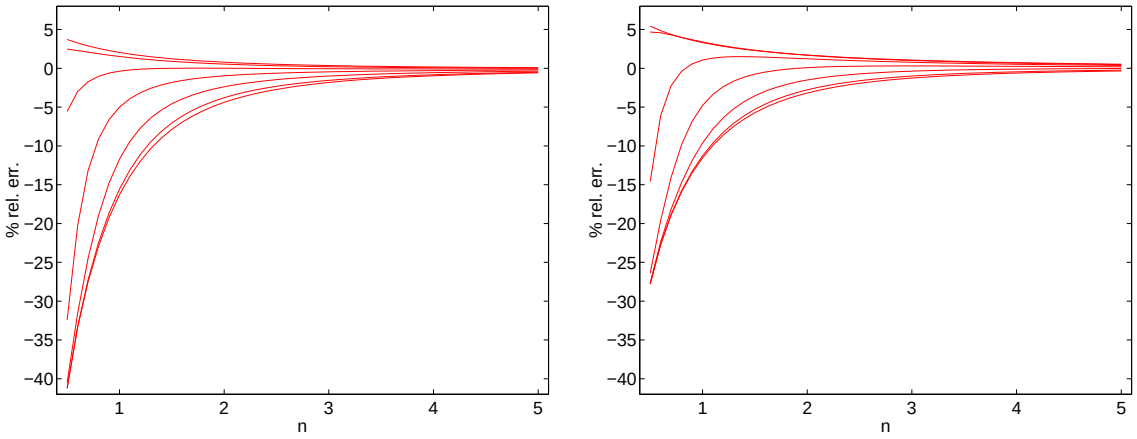

Figure 3.3: Left panel: The relative percentage error of the SPA to $F_{t''}(t; n, 0, \theta)$ as a function of the degrees of freedom parameter n, where t is chosen such that $F_{t''}(t; n, \mu, \theta) = 0.95$, $\mu = 0$, and $\theta = 0, 1, \ldots, 6$ (from bottom to top). **Right panel**: Same, but using $\mu = 10$.

3.5 Empirical Application

We illustrate the use of the new computational method in the context of modelling the daily returns on the financial index NASDAQ which, like essentially all returns on financial assets observed at weekly or higher frequencies, are characterized by high excess kurtosis relative to the normal distribution ("fat tails"), mild skewness, and strong volatility clustering. The interest in modelling the returns on financial assets has grown enormously over the past decade. From the viewpoint of risk managers in financial institutions, short-term out-of-sample quantile prediction is of extreme importance (most notably the accurate estimation of the Value-at-Risk measure, which is currently the most popular and important measure of risk used in risk management; see Dowd (2005) and Kuester et al. (2006) for detailed overviews of the rampantly growing literature in this field).

Because the (central) Student's t distribution can assume fat–tailed behavior at least as extreme as the Cauchy and also nests (in a limiting sense) the normal, it has become the de facto standard choice of innovations distribution for both unconditional (in which the returns are treated as being i.i.d.) and conditional (in which a stochastic process for the time-varying location and, more importantly, scale term, is used) modelling of financial returns data (see, e.g., Palm, 1996, for a detailed overview). Numerous empirical papers have demonstrated that the Student's t can accommodate the fat–tailed nature of the data very well, but, obviously, not the skewness. To account for this, several asymmetric extensions of the Student's t distribution have been proposed and demonstrated to be highly effective in a risk forecasting context. These include the singly noncentral t, as first advocated in this context by Harvey and Siddique (1999).

In the conditional modelling case, by far the most popular method for capturing the strong stochastic volatility component inherent in financial returns is use of a GARCH–type model, which we briefly discuss. Denote the return at time t by R_t, $t = 1, \ldots, T$, which we model as $R_t = a_0 + Z_t \sigma_t$, where the Z_t are assumed to be i.i.d.. To capture the evolution of the scale parameter σ_t, we use the popular Asymmetric Power ARCH, or A-PARCH, model proposed by Ding et al. (1993), given by

$$\sigma_t^\delta = c_0 + \sum_{i=1}^{r} c_i \left(|\epsilon_{t-i}| - \gamma_i \epsilon_{t-i} \right)^\delta + \sum_{i=1}^{s} d_i \sigma_{t-i}^\delta, \quad \epsilon_t = Z_t \sigma_t, \tag{3.20}$$

with $c_i > 0$, $d_i \geq 0$, $\delta > 0$, and $|\gamma_i| < 1$. As detailed in Mittnik et al. (1998), correct estimation of (3.20) and handling of the initial $\max(r, s)$ values of σ_t and ϵ_t necessitate knowing $\mathbb{E}\left[\left(|Z| - \gamma Z \right)^\delta \right]$. This can be effectively and nearly instantly approximated by simple numeric integration when using the SPA for the singly or doubly noncentral t.

In our (and most GARCH) applications, choosing $r = s = 1$ and setting $\delta = 1$ suffices. The model (3.20), when coupled with the singly noncentral t distribution for the Z_t, i.e., $Z_t \overset{\text{iid}}{\sim} t'(n, \mu)$,

has been shown to be very successful in a risk prediction context. We thus require

$$\kappa := \mathbb{E}\big[\,(|Z| - \gamma Z)\,\big] = \mathbb{E}\big[\,|Z|\,\big] - \gamma\mathbb{E}\big[Z\big], \quad Z \sim t'(n, \mu),$$

which can be analytically determined: an expression for the first moment of $T \sim t''(n, \mu, \theta)$ is given in (3.12) above, while one for $\mathbb{E}[|T|]$ does not appear to have been considered in the literature. Straightforward calculation shows that

$$\mathbb{E}(|T|^r) = n^{r/2}\frac{\Gamma\left((n-r)/2\right)\Gamma\left((r+1)/2\right)}{\Gamma\left(n/2\right)\Gamma\left(1/2\right)}{}_1F_1\left(\frac{r}{2}, \frac{n}{2}, -\frac{\theta}{2}\right){}_1F_1\left(-\frac{r}{2}, \frac{1}{2}, -\frac{\mu^2}{2}\right), \quad r < n. \quad (3.21)$$

The confluent hypergeometric functions ${}_1F_1$ appearing in (3.12) and (3.21) could be quickly approximated using the approximation given in Butler and Wood (2002). Based on (3.12) and (3.21), κ can be computed.

For the daily NASDAQ returns from June 1993 to June 2001 (about 2000 observations), the maximum likelihood estimates using the exact singly noncentral t density are

$$\hat{a}_0 = 0.185\,(0.046), \quad \hat{c}_0 = 0.0187\,(0.0057), \quad \hat{c}_1 = 0.112\,(0.016), \quad \hat{d}_1 = 0.884\,(0.017),$$
$$\hat{\gamma}_1 = 0.283\,(0.078), \quad \hat{n} = 9.15\,(1.7), \quad \hat{\mu} = -0.0870\,(0.049).$$

As is common in this and related statistical applications (see, e.g., the detailed discussion in Morgan, 2000, Chapters 3 and 4), the approximate standard errors, given in parentheses, have been obtained from the numerical Hessian, evaluated at the estimates. The estimation required about 60 iterations using a quasi-Newton method for multivariate optimization; this corresponds to about 550 evaluations of the likelihood, each of which required computing the pdf for each of the 2000 data points, i.e., over one million evaluations of the singly noncentral t density are required. Use of the SPA with renormalization via numeric integration yields

$$\hat{a}_0 = 0.188\,(0.044), \quad \hat{c}_0 = 0.0191\,(0.0055), \quad \hat{c}_1 = 0.112\,(0.016), \quad \hat{d}_1 = 0.883\,(0.017),$$
$$\hat{\gamma}_1 = 0.281\,(0.078), \quad \hat{n} = 8.97\,(1.7), \quad \hat{\mu} = -0.0919\,(0.047).$$

The point estimates based on the SPA barely differ from those using the exact pdf calculation, and their differences are negligible with respect to the reported standard errors. Using the SPA with the response surface approximation to the constant of integration yields

$$\hat{a}_0 = 0.191\,(0.034), \quad \hat{c}_0 = 0.0143\,(0.0054), \quad \hat{c}_1 = 0.104\,(0.015), \quad \hat{d}_1 = 0.895\,(0.016),$$
$$\hat{\gamma}_1 = 0.279\,(0.076), \quad \hat{n} = 8.73\,(1.6), \quad \hat{\mu} = -0.0947\,(0.039),$$

which are also not significantly different from either the values based on the exact pdf or the renormalized-via-integration SPA.

With regard to the (statistically insignificant and practically negligible) parameter difference based on the renormalized SPA and true density, it is understood in such empirical finance applications that the model innovations are certainly not really distributed as a noncentral t; it is just an excellent parametric approximation because of its flexible skewness and kurtosis (but does offer the plausible interpretation as a continuous mixture of normals; see Praetz, 1972). As such, in this context, the fact that the renormalized SPA is not exact is fully irrelevant; being a proper density in its own right, its use is justified just as well as that of the true density.

To illustrate the importance of renormalization, use of the SPA without renormalization gave highly different parameters; for example, $\hat{n} = 1.18$, which would misleadingly indicate that the GARCH innovations could follow a Cauchy distribution and thus highly overestimate the risk of a long or short position in the asset. Use of the adjusted pdf (3.9) performed between the two extremes: the estimated degrees of freedom in this case is $\hat{n} = 3.90$, which is too low. The explanation of this disappointing performance is that the amount of skewness in this data set (and most financial return series) is not great, so that the distribution is very close to being the regular Student's t, for which the renormalized SPA is *exact*.

It is well-known in the empirical finance literature that the distributional assumption on the conditional innovation sequence of a GARCH model is far more decisive for the quality of in-sample fit and out-of-sample forecasting ability than the form of the GARCH recursion for the conditional volatility (see, e.g., the extensive results in Mittnik and Paolella, 2000, and Bao et al., 2003). As such, it might be expected that use of the doubly noncentral t distribution (with the additional parameter θ viewed as another shape parameter) in model (3.20) would provide a better description of the data generating process than use of the singly noncentral t. To the best of the authors' knowledge, this has never been attempted, owing, in all likelihood, to the computational complexity of the t'' pdf. If, however, the exact pdf is replaced by its SPA, the estimation of such a model becomes feasible, requiring in fact no more time than with the singly noncentral t distribution.

For the NASDAQ data set, the estimate of $\hat{\theta}$ was insignificantly different from zero, based on the asymptotic normality of the MLE (see Straumann, 2005, Chapter 6.3). Nevertheless, given the plethora of different kinds of financial data sets and their well-reported distributional asymmetries (see, e.g., Cappuccio et al., 2004, Premaratne and Bera, 2005, Lisi, 2005, and the references therein), it seems likely that the additional flexibility of the doubly noncentral t could be of potentially great value in this modelling context.

Appendix 3.A Proof of the Uniqueness of the Saddlepoint

We first prove that the discriminant $m < 0$, i.e., all three roots are real. The discriminant can be written as

$$\begin{aligned} m = &- \frac{n}{108\left(y_1^2+n\right)^6}\left(-8y_1^2\mu^2\theta^2 n + 4\theta^3 n^2 + 4y_1^4\mu^4\theta\right) - \\ &- \frac{n}{108\left(y_1^2+n\right)^6}\left(ny_1^6\mu^2 + 12n^2y_1^4\theta + 2n^2y_1^4\mu^2 + 12n^2y_1^2\theta^2 + 24n^3y_1^2\theta + y_1^2\mu^2n^3 + \right. \\ &\left. + 4n^2y_1^6 + 12n^3y_1^4 + 12n^4y_1^2 + 12\theta^2n^3 + 12\theta n^4 + 4n^5 + 20ny_1^4\mu^2\theta + 20y_1^2\mu^2\theta n^2\right). \end{aligned}$$

As $n > 0$ and $\theta \geq 0$, all terms in the second parentheses are positive. The first term in parentheses can be factored as $-8y_1^2\mu^2\theta^2 n + 4\theta^3 n^2 + 4y_1^4\mu^4\theta = 4\theta(\theta n - y_1^2\mu^2)^2$, so that m cannot be positive, whence all three roots are real and, as shown in Butler and Paolella (2002), ordered as $z_2 < z_3 < z_1$. Furthermore, as $a_3 > 0$ and $\operatorname{sgn}(a_2) = -\operatorname{sgn}(y_1\mu) = -\operatorname{sgn}(a_0)$, the coefficients of $s(y_2)$ can have at most two variations in sign, and thus, from Descartes' rule of signs, $s(y_2)$ can have at most two positive roots. Similarly, the coefficients of $s(-y_2)$ can have at most two variations in sign, and therefore $s(y_2)$ has at least one positive root. Consequently,

$$z_2 < 0 \tag{3.22}$$

and $z_1 > 0$. Also, we require that $\hat{t}_2 < 1/2$, or, plugging in $\hat{t}_1$ and $\hat{t}_2$ from (3.6),

$$\hat{y}_2 > \frac{y_1\mu}{y_1^2+n} =: y_2^*. \tag{3.23}$$

Next observe that, as $a_3 > 0$ and $m < 0$,

$$s(y_2)\begin{cases} < 0, & y_2 < z_2, \\ > 0, & z_2 < y_2 < z_3, \\ < 0, & z_3 < y_2 < z_1, \\ > 0, & y_2 > z_1, \end{cases} \tag{3.24}$$

as well as that $s(0) = y_1\mu n$ and $s(y_2^*) = -y_1\mu n\theta/\left(y_1^2+n\right)$. If $y_2^* > 0$, then $s(y_2^*) < 0$, and it follows from (3.22) and (3.24) that $z_3 < y_2^* < z_1$, so that the saddlepoint solution must be z_1 from (3.23). If $y_2^* \leq 0$, then $s(0) \leq 0$ and, from (3.22) and (3.24), $z_3 \leq 0 < z_1$, implying that the saddlepoint solution must be $z_1 > 0$.

Manuscript 4

CHICAGO: A Fast and Accurate Method for Portfolio Risk Calculation

4.1 Introduction

The Value at Risk and Expected Shortfall, or VaR and ES, respectively, of a portfolio of assets have become key figures in risk management. Accurately forecasting these measures has, however, proved challenging, because returns on financial assets observed at weekly or higher frequencies exhibit excess kurtosis and mild skewness, and because the assets are not independent of each other.

The approach of fitting a univariate GARCH-type model directly to the time series of portfolio returns is effective, but bears the problem that the entire model must be re-estimated every time the composition of the portfolio changes. The conventional solution is to model the joint evolution of the assets by means of multivariate GARCH, or MGARCH models. Many different MGARCH models have been considered in the literature, with the common goal of reducing the dimensionality of the most general such model, the VEC(p,q) model of Bollerslev et al. (1988); an excellent overview is given in Bauwens et al. (2006). Among these models, the constant conditional correlation (CCC) model of Bollerslev (1990) and the dynamic conditional correlation (DCC) model of Engle (2002) bear the advantage that there exist two-step procedures for their estimation, thus effectively reducing the multivariate problem to a set of univariate ones. As such, estimation of high-dimensional models becomes feasible, albeit at the expense of flexibility. More complex models that offer a more general covariance specification typically do not allow such a simplification. Examples are the BEKK model of Engle and Kroner (1995), and the Factor-GARCH and Generalized Orthogonal (GO) GARCH models of Engle et al. (1990) and van der Weide (2002), respectively, the latter two being nested in the former. Therefore, inference in high-dimensional problems is practically infeasible, especially in a non-Gaussian setup.

This manuscript will demonstrate a two-step procedure for estimating the GO-GARCH model of van der Weide (2002) with non-Gaussian innovations. Our main tool to achieve this is independent component analysis (ICA), details on which will be given below; in order to motivate the name of our method — CHICAGO, or *C*onditionally *H*eteroscedastic *I*ndependent *C*omponent *A*nalysis of *G*eneralized *O*rthogonal GARCH models, suffice it for now that the specific ICA algorithm employed in this paper maximizes the conditional heteroscedasticity of the estimated components. As for the CCC and DCC models, ICA allows us to decompose the problem into a set of readily estimable univariate models, while at the same time maintaining sufficient flexibility in the specification of the co-evolution of the assets. The only missing link to an efficient calculation of portfolio Value at Risk is then the evaluation of tail probabilities for a portfolio of assets. We achieve this by modelling the conditional distribution of the individual assets as generalized hyperbolic (GHyp); in this sense, our approach is similar to that of Chen et al. (2006), who employ the GHyp distribution and ICA in a non-parametric setting. The present paper replaces the Fast Fourier Transform used by the latter authors to obtain the required distribution of convolutions of independent GHyp random variables by a saddlepoint approximation.

The remainder of this manuscript is organized as follows. Section 4.2 introduces the GO-

GARCH model, and shows how independent component analysis can be used to estimate it in two steps. Section 4.3 details the univariate factor specification. Section 4.4 derives the saddlepoint approximation for convolutions of independent GHyp distributions. Section 4.5 details the results of a backtesting exercise. Section 4.6 concludes.

4.2 The GO-GARCH Model

4.2.1 Model Specification

Consider a set of d financial assets, with associated return vector $\mathbf{r}_t$, $t \in \{1 \dots, T\}$. In the GO-GARCH model of van der Weide (2002), the innovations $\mathbf{u}_t$ are modelled as linear combinations of d unobserved factors $\mathbf{f}_t$:

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{u}_t \tag{4.1}$$

$$\mathbf{u}_t = \mathbf{A}\mathbf{f}_t, \tag{4.2}$$

for some mixing matrix $\mathbf{A}$ that is invertible and constant over time. The unobserved factors are assumed to be independent of each other, and to have unit unconditional variance. Note that the latter is an identifying restriction; any scale factors can simply been absorbed into the mixing matrix. It follows that the unconditional covariance matrix of the returns is given by

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{u}_t \mathbf{u}_t'\right] = \mathbf{A}\mathbf{A}'.$$

If one assumes a GARCH(1,1) process for each factor $\{f_{it}\}$, i.e, $\mathbf{f}_t \sim (\mathbf{0}, \mathbf{H}_t)$, where

$$\mathbf{H}_t = \boldsymbol{\Omega} + \sum_{k=1}^{d} \alpha_k \mathbf{e}_k \mathbf{e}_k' \mathbf{H}_{t-1} \mathbf{e}_k \mathbf{e}_k' + \sum_{i=1}^{d} \beta_k \mathbf{e}_k \mathbf{e}_k' \mathbf{f}_{t-1} \mathbf{f}_{t-1}' \mathbf{e}_k \mathbf{e}_k',$$

$\boldsymbol{\Omega} = \sum_{k=1}^{d} (1 - \alpha_k - \beta_k) \mathbf{e}_k \mathbf{e}_k'$, and $\mathbf{e}_k$ is a $d \times 1$ vector with kth element 1 and zeros everywhere else, then the conditional covariance of the return series becomes

$$\boldsymbol{\Sigma}_t = \mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \sum_{k=1}^{d} \alpha_k \lambda_k w_k' \boldsymbol{\Sigma}_{t-1} w_k \lambda_k' + \sum_{i=1}^{d} \beta_k \lambda_k w_k' \mathbf{u}_{t-1} \mathbf{u}_{t-1}' w_k \lambda_k',$$

where $\lambda_k = \mathbf{A}\mathbf{e}_k$ and $w_k = (\mathbf{A}^{-1})'\mathbf{e}_k$. This is a special case of the factor GARCH model of Engle et al. (1990), and of the BEKK model of Engle and Kroner (1995). In contrast to the latter, however, the GO-GARCH model does not necessarily require that the underlying factors obey a GARCH structure, so that more flexible dynamic specifications such as, e.g., the Asymmetric Power ARCH, or A-PARCH, model of Ding et al. (1993), can be employed. The Full Factor Model of Vrontos et al. (2003) is based on a similar idea, but uses a triangular mixing matrix.

Using a polar decomposition, the mixing matrix $\mathbf{A}$ can be uniquely factorized into a symmetric positive definite matrix $\mathbf{\Sigma}^{1/2}$ and an orthogonal matrix $\mathbf{U}$:

$$\mathbf{A} = \mathbf{\Sigma}^{1/2}\mathbf{U}, \tag{4.3}$$

where $\mathbf{\Sigma}^{1/2}$ is the symmetric positive definite square root of the unconditional covariance matrix $\mathbf{\Sigma}$, so that

$$\mathbf{A}\mathbf{A}' = \mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{U}'\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}, \tag{4.4}$$

because $\mathbf{U}\mathbf{U}' = \mathbf{I}$. Note that, as the factors $\mathbf{f}_t$ are unobserved, their order is not identifiable. Hence, the order of the columns of $\mathbf{U}$ can be chosen arbitrarily to make the determinant positive, so that $\mathbf{U}$ is a rotation matrix. As a rotation matrix, $\mathbf{U}$ can be decomposed as the product of $\binom{d}{2}$ basic rotation matrices $\mathbf{R}_i(\theta_i)$, where each $\mathbf{R}_i$ is a rotation of angle θ_i in the plane spanned by one pair of axes in $\mathbb{R}^d$ (see van der Weide, 2002, Lemma 3). Thus, $\mathbf{U}$ can be fully parameterized in terms of the Euler angles θ_i. For example, in the $d = 2$ case,

$$\mathbf{U} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

The angles θ_i can be restricted to the interval $0 \leq \theta_i < \pi/2$, because the signs and order of the factors cannot be identified.

4.2.2 Estimation of the Mixing Matrix

4.2.2.1 Estimation Strategies

If one assumes a GARCH(1,1) structure for each factor, then the GO-GARCH model contains $d^2 + 3d$ parameters. The difficulties associated with jointly estimating a parameter vector of this dimension make it desirable to have a two-step procedure available, in which the mixing matrix is estimated first, thus effectively reducing the problem to a set of d univariate problems. For any two-step algorithm, it is expedient to exploit the decomposition in (4.3), because the $d(d+1)/2$ free parameters in $\mathbf{\Sigma}^{1/2}$ can be consistently estimated from the (unconditional) sample covariance. As such, we will only consider the whitened and demeaned data

$$\mathbf{z}_t = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\mathbf{u}}_t, \tag{4.5}$$

where $\hat{\mathbf{u}}_t \equiv (\mathbf{r}_t - \hat{\boldsymbol{\mu}}_t)$, and we assume that the time-dependent mean can be consistently estimated by ordinary least squares (Note that this requirement precludes GARCH-in-mean-type models.). The unconditional covariance matrix could be estimated from the sample covariance of the OLS residuals; however, though unbiased, this simple estimator may be inadequate for the high-dimensional data sets for which the method is designed. As such, we propose to use the shrinkage estimator of Ledoit and Wolf (2003), which, apart from being more efficient, guarantees positive

definiteness of $\hat{\mathbf{\Sigma}}$.

The $d(d-1)/2$ parameters in $\mathbf{U}$, on the other hand, cannot be estimated on the basis of unconditional information alone, because from (4.4), any orthogonal matrix $\mathbf{U}^*$ gives rise to the same unconditional covariance matrix $\mathbf{\Sigma}$. Thus, any estimation method must utilize conditional information. One such procedure is given in Boswijk and van der Weide (2006), where an estimator $\hat{\mathbf{U}}_{BW}$ is derived as the eigenvector matrix of the symmetric matrix $\hat{\mathbf{B}}$ solving the nonlinear least squares problem

$$\hat{\mathbf{B}} = \text{argmin}_{\mathbf{B}:\mathbf{B}=\mathbf{B}'} \frac{1}{T} \sum_{t=1}^{T} \text{tr}\left(\left[\mathbf{z}_t \mathbf{z}_t' - \mathbf{I}_d - \mathbf{B}\left(\mathbf{z}_{t-1}\mathbf{z}_{t-1}' - \mathbf{I}_d \right) \mathbf{B} \right]^2 \right).$$

We propose here to estimate $\mathbf{U}$ by independent component analysis (ICA), which we briefly outline next. Details can be found in the monograph by Hyvärinen et al. (2001).

The basic ICA model assumes that a d-dimensional random vector $\mathbf{u}_t \equiv [u_{1t}, \ldots, u_{dt}]'$ is observed. The u_{it} are linear combinations of d independent random variables f_{it},

$$u_{it} = \sum_{j=1}^{d} a_{ij} f_{jt},$$

or, in matrix form,

$$\mathbf{u}_t = \mathbf{A}\mathbf{f}_t.$$

The aim is to estimate both $\mathbf{f}_t$ and $\mathbf{A}$, i.e., to find a matrix $\mathbf{W} \equiv \mathbf{A}^{-1}$ such that $\mathbf{y}_t \equiv \mathbf{W}\mathbf{u}_t$ are independent. The naive approach of taking $\mathbf{W} = \hat{\mathbf{\Sigma}}^{-1/2}$ produces uncorrelated components, but yields independent components only up to an orthogonal transformation: With $\mathbf{U}^*$ orthogonal,

$$\mathbb{E}[\mathbf{U}^* \mathbf{y}_t \mathbf{y}_t' \mathbf{U}^{*\prime}] = \mathbf{U}^* \mathbf{U}^{*\prime} = \mathbf{I},$$

i.e., $\mathbf{U}^* \mathbf{y}_t$ is also uncorrelated, but not necessarily independent, unless the the data are i.i.d. multivariate Gaussian. In terms of the decomposition (4.3), the orthogonal matrix $\mathbf{U}$ remains to be estimated, which requires information beyond that contained in the unconditional covariance matrix. Which particular additional information to use depends on the problem at hand; for example, with non-Gaussian data, the central limit theorem can be exploited: the distribution of a normalized sum of independent random variables with finite second moments converges to a Gaussian. Conversely, a linear combination $\mathbf{w}'\mathbf{u}_t = \mathbf{w}'\mathbf{A}\mathbf{f}_t =: \mathbf{q}'\mathbf{f}_t$ will be "least Gaussian" if $\mathbf{q}$ has one element equal to 1 and all others zero. The degree of non-Gaussianity can be measured by negentropy, which for a random variable X with density f_X is defined as

$$J(f_X) = S(\phi_X) - S(f_X),$$

where $S(f) = -\int f(x) \log f(x) \mathrm{d}x$ is the differential entropy of f, and ϕ_X denotes the density

of a Gaussian random variable with the same variance as X. The aim is then to find $\mathbf{w}$ such that the negentropy (or an approximation thereof, as the density is typically unknown) of $\mathbf{w}'\mathbf{r}_t$ is maximized. The FastICA algorithm of Hyvärinen (1999) achieves this with cubic convergence.

For time series data, exploiting the time structure of the data set to identify the independent components appears more natural. Financial returns data typically exhibit very strong GARCH-effects, and it suggests itself to rely on these as additional information. It is well known that the GARCH-effects present in the sum of two (or more) series are weaker than those in the individual series themselves. It is therefore possible to separate the independent components by maximizing the autocorrelation of the squared returns. Hyvärinen et al. (2001, p. 349) devise a fixed-point algorithm based on cross cumulants that achieves this with cubic convergence. Given prewhitened data $\mathbf{z}_t$ as in (4.5), the algorithm starts with $\mathbf{U}_n = \mathbf{I}$ and iterates

$$\begin{aligned}\mathbf{U}_{\text{temp}} &= \mathbf{z}[\mathbf{z}'\mathbf{U}_n \odot \mathbf{z}'_-\mathbf{U}_n \odot \mathbf{z}'_-\mathbf{U}_n]/T + \mathbf{z}_-[\mathbf{z}'_-\mathbf{U}_n \odot \mathbf{z}'\mathbf{U}_n \odot \mathbf{z}'\mathbf{U}_n]/T - 2\mathbf{U}_n - 4\bar{\mathbf{C}}\mathbf{U}_n\mathbf{D}_n \\ \mathbf{U}_{n+1} &= (\mathbf{U}_{\text{temp}}\mathbf{U}'_{\text{temp}})^{-1/2}\mathbf{U}_{\text{temp}}\end{aligned}$$

where $\mathbf{z} = [\mathbf{z}_2, \ldots, \mathbf{z}_T]$, $\mathbf{z}_- = [\mathbf{z}_1, \ldots, \mathbf{z}_{T-1}]$, $\bar{\mathbf{C}} = (\mathbf{z}\mathbf{z}'_- + \mathbf{z}_-\mathbf{z}')/(2T)$, $\mathbf{D}_n = \text{diag}\left(\text{vecd}\left(\mathbf{U}'_n\bar{\mathbf{C}}\mathbf{U}_n\right)\right)$, and $\odot$ denotes the Hadamard product. The iteration stops when $1-c < \epsilon$, where c is the minimum over the absolute values of the diagonal elements of $\mathbf{U}'_{n+1}\mathbf{U}_n$, and ϵ is a suitable convergence threshold (we use 10^{-12}). In the rare cases that the algorithm fails to converge, one may fall back to the negentropy-based FastICA algorithm.

Figures 4.1 and 4.2 illustrate the technique. From top to bottom, the rows of each Figure plot the original components $\mathbf{f}_t$, the mixed components $\mathbf{A}\mathbf{f}_t$, and the estimated components. It is apparent from the graphs that ICA is able to restore the original components, except for their signs and order, which are not identifiable.

4.2.2.2 Performance Comparison

Thus, there exist (at least) three different estimators of the rotation matrix $\mathbf{U}$: the maximum likelihood estimator (MLE), the estimator of Boswijk and van der Weide (BW), and independent component analysis (ICA). A simulation study is useful to compare their relative performance, and we detail the results of such an experiment next. We consider a bivariate model, which bears the advantage that the results can be condensed into a single statistic, namely, the estimated rotation angle $\hat{\theta}$ of the rotation matrix $\mathbf{U}$. To keep matters simple, we use a GARCH(1,1) model with Gaussian innovations for each factor, with parameters $(\alpha_1, \beta_1) = (0.09, 0.9)$ and $(\alpha_2, \beta_2) = (0.04, 0.95)$. Note that for both factors, the parameters are close to the stationarity border ($\alpha_i + \beta_i = 0.99$ for both factors), which closely mimics the characteristics typically found in actual data. Because our interest centers on the rotation angle θ, the unconditional covariance matrix was set to the identity matrix, and the length of each of the 1,000 simulated samples is $T = 800$. We report the root mean squared error (RMSE) and BIAS of the estimated angle $\hat{\theta}$

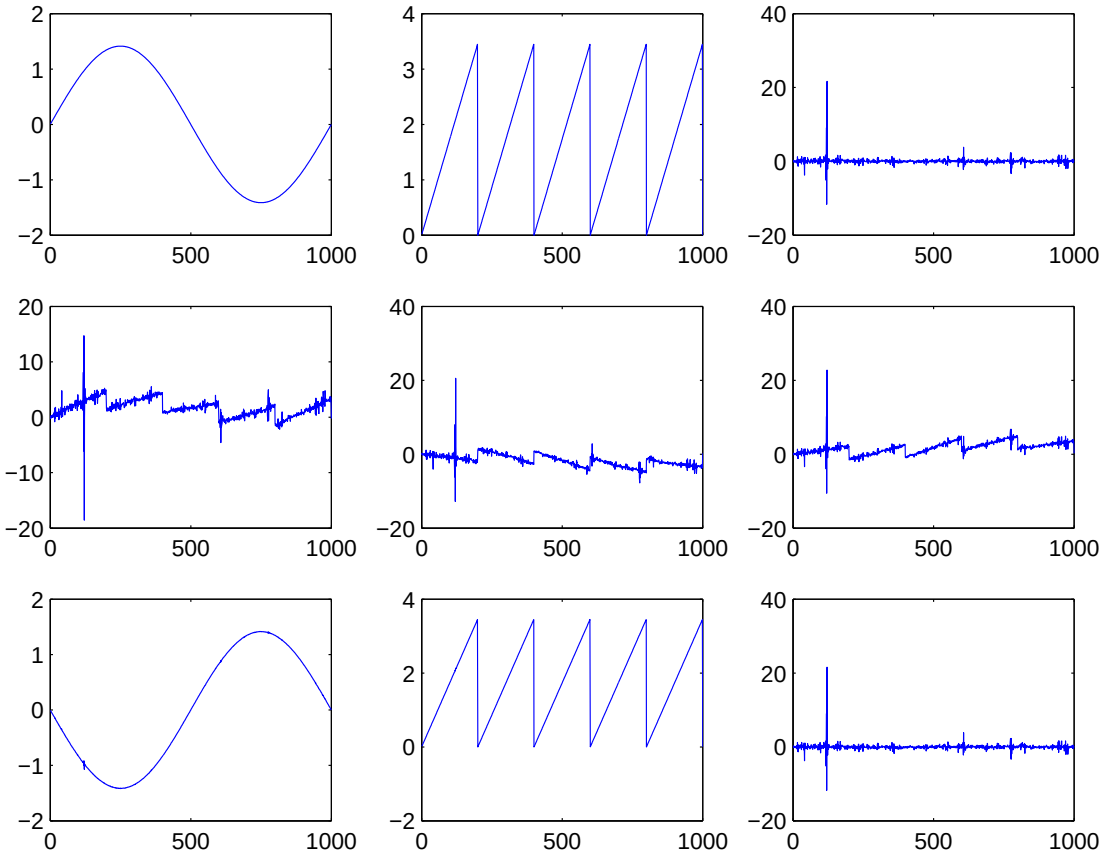

Figure 4.1: Example for ICA. Top row: original signals. Middle row: mixed signals. Bottom row: estimated independent components.

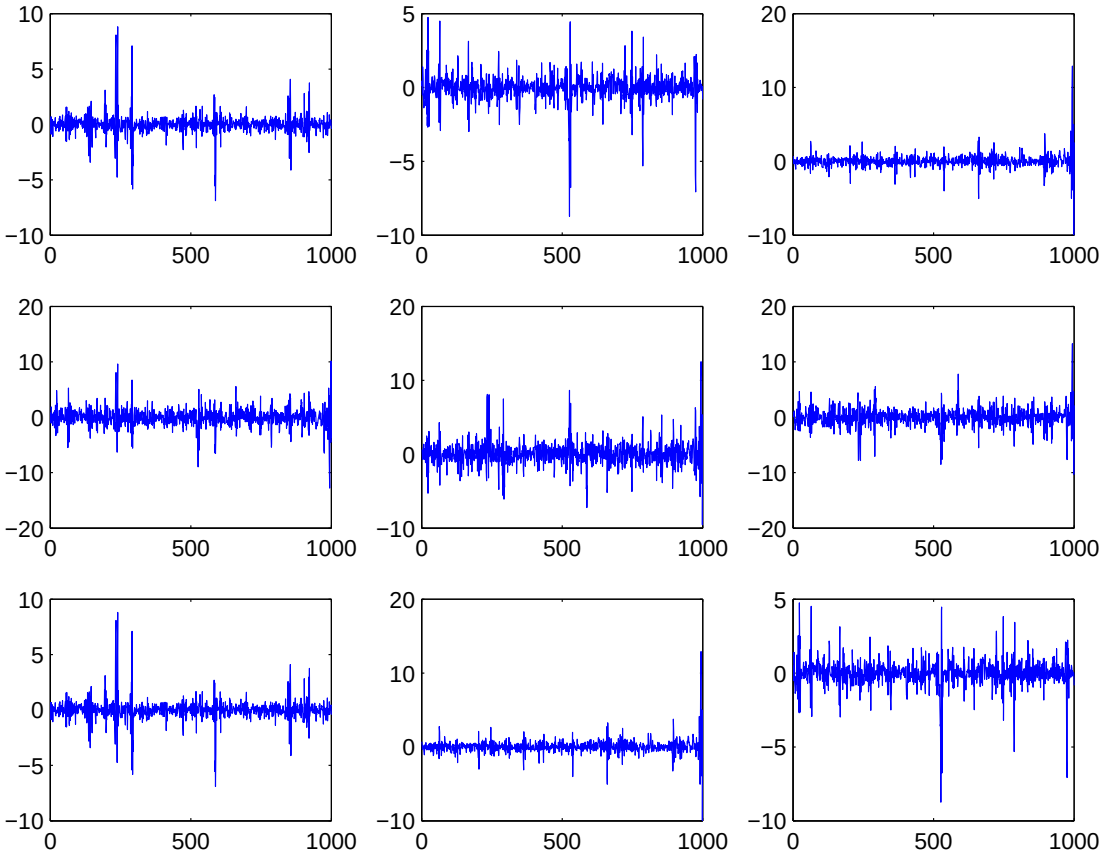

Figure 4.2: Example for ICA. Top row: original signals. Middle row: mixed signals. Bottom row: estimated independent components.

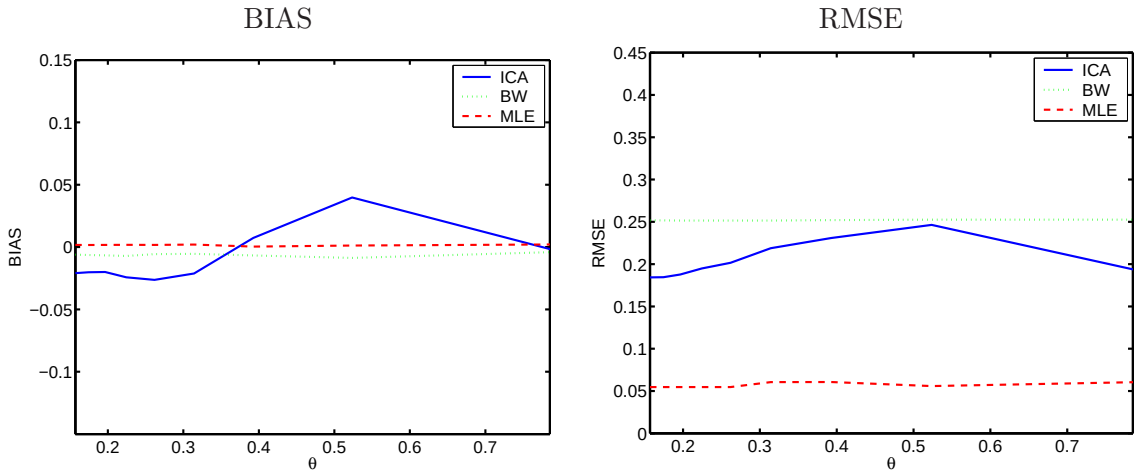

Figure 4.3: Performance Comparison of MLE, ICA, and Boswijk and van der Weide (BW).

for each of the three estimation methods. The MLE — like the other two estimators — was computed from demeaned and whitened data, rather than jointly estimating the unconditional covariance matrix $\mathbf{\Sigma}$ with the remaining parameters.

A graphical representation of the results for different values of the true rotation angle θ is given in Figure 4.3. For each estimated rotation matrix $\hat{\mathbf{U}} = (u_{ij})$, the rotation angle was computed as $\hat{\theta} = \cos^{-1}(u_{11})\text{sgn}(u_{12}) + k\pi/2$, where k is an integer minimizing the distance between $\hat{\theta}$ and the true value, θ (note that the rotation angle is only identifiable up to multiples of $\pi/2$). For the parameter constellation under investigation, the ICA estimator displays a slight bias, whereas both the MLE and BW estimators are virtually unbiased. In terms of RMSE, the MLE is favored over the alternative estimators, and the ICA estimator has a slight yet consistent advantage over the estimator of Boswijk and van der Weide. However, it should be borne in mind that the BW and ICA estimators use only information contained in the autocorrelation of the squared returns, whereas the MLE relies on the exact specification of the factor dynamics; as such, the former two estimators are vastly more robust with respect to the assumptions on the evolution of the univariate factors.

Another important remark concerns the computation time required by each estimator: the MLE, for the sample size at hand, took on average 8.87 seconds to compute, the estimator of Boswijk and van der Weide 1.68 seconds, and the ICA estimator 0.03 seconds, rendering the latter method 297 and 56 times faster, respectively, than the former two. Lastly, the ICA method converges reliably even for high-dimensional time series, without any of the convergence problems associated with maximum likelihood estimation in such cases. Taking all aspects (accuracy, robustness, computational reliability, and speed) into account, independent component analysis is a very useful tool for estimating GO-GARCH models.

4.3 Univariate Factor Specifiation

4.3.1 Marginal Distribution of Assets

We assume a generalized hyperbolic law for the conditional distribution of each factor. The GHyp is an extremely flexible asymmetric and fat–tailed distribution which nests a large number of distributions that are popular in the empirical modelling of asset returns. These include the Student's t, which is ubiquitous in such applications (see, e.g., the survey article of Palm, 1996, and the references therein); the Laplace (see, e.g., Granger and Ding, 1995; Mittnik et al., 1998; Haas et al., 2006), normal inverse Gaussian (NIG), and hyperbolic (see Küchler et al., 1999; Prause, 1999). While the GHyp offers an excellent empirical fit to financial data, its use can also be motivated from a theoretical point of view. For example, Barndorff-Nielsen (1998) discusses a stochastic volatility process with normal inverse Gaussian marginals. Also, Reimann (2005) has demonstrated that a purely economic model for asset return generation results in empirical behavior strikingly similar to a hyperbolic distribution.

Value at Risk for portfolios of multivariate generalized hyperbolic assets has been considered in Bauer (2000) and Sadefo-Kamdem (2006), and, for the important special case of the multivariate NIG, in Aas et al. (2006). In our model, the factors conditionally obey a different, non-elliptical multivariate generalization of the GHyp, in which each component can be expressed as a linear combination of independent univariate GHyp random variables. This distribution has been analyzed in Schmidt et al. (2006), where it was found to be adequate for returns data.

Specifically, let $\lambda \in \mathbb{R}$, $\omega > 0$, $-1 < \rho < 1$, $\mu \in \mathbb{R}$ and $\sigma > 0$. Then random variable X follows a generalized hyperbolic density, written $X \sim \mathrm{GHyp}(\lambda, \omega, \rho, \sigma, \mu)$, if its density is given by

$$f_X(x; \lambda, \omega, \rho, \sigma, \mu) = \frac{\omega^\lambda \bar{y}^{\lambda-\frac{1}{2}}}{\sqrt{2\pi}\bar{\alpha}^{\lambda-\frac{1}{2}}\sigma K_\lambda(\omega)} K_{\lambda-\frac{1}{2}}\left(\bar{\alpha}\bar{y}\right) e^{\rho\bar{\alpha}z}, \tag{4.6}$$

where $z \equiv \frac{x-\mu}{\sigma}$, $\bar{\alpha} \equiv \omega(1-\rho^2)^{-1/2}$, $\bar{y} \equiv \sqrt{1+z^2}$, and $K_\nu(x)$ is the the modified Bessel function of the third kind with index ν, defined as

$$K_\nu(x) = \frac{1}{2}\int_0^\infty t^{\nu-1} e^{-\frac{1}{2}x(t+t^{-1})}\,\mathrm{d}t.$$

Reliable numeric routines exist, for example in Matlab, for computing $K_v(x)$. The parameters of the generalized hyperbolic have the following interpretation: μ and σ are genuine location and scale parameters, respectively, while λ, ω and ρ are location–scale invariant. Parameter ω controls the tail thickness, and ρ is a measure of the skewness. We will consider the special cases obtained by letting $\lambda = -\frac{1}{2}$ (normal inverse Gaussian, or NIG), and $\lambda = 1$ (hyperbolic). We found that these special cases still offer enough flexibility to capture the excess kurtosis and skewness present in financial returns data: Figures 4.4 and 4.5 show the excellent fit offered by the standardized NIG and hyperbolic distributions, respectively, for two independent components identified from

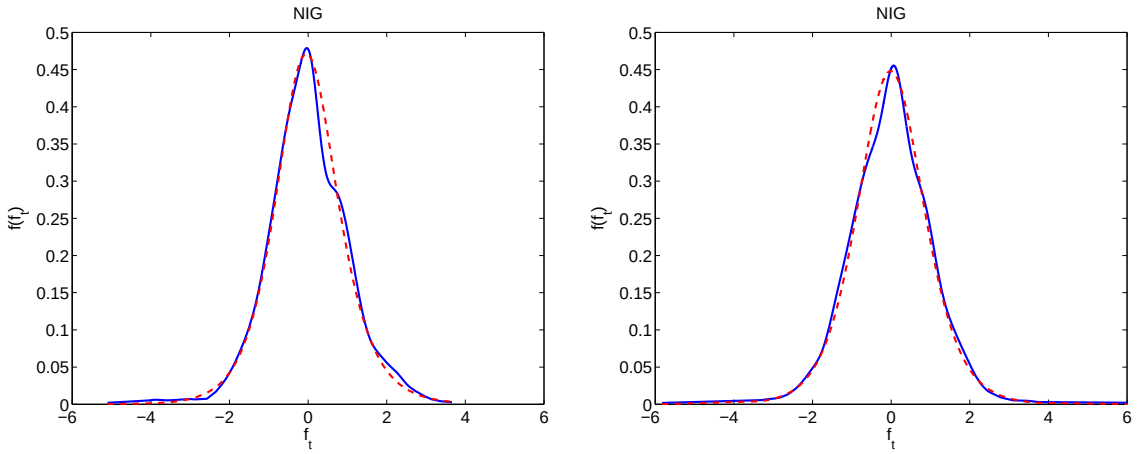

Figure 4.4: Kernel density (solid) of filtered residuals and fitted NIG densities (dashes) for two ICs.

the Dow Jones data set considered in Section 4.5.

The expected value and variance of the GHyp distribution are given by

$$\mathbb{E}[X] = \mu + \sigma \frac{\rho}{\sqrt{1-\rho^2}} k_1(\omega) \tag{4.7}$$

and

$$\mathbb{V}(X) = \sigma^2 \left[\omega^{-1} k_1(\omega) + \frac{\rho^2}{1-\rho^2} k_2(\omega) \right],$$

respectively, where $k_1(\omega) \equiv K_{\lambda+1}(\omega)/K_\lambda(\omega)$ and $k_2(\omega) \equiv [K_\lambda(\omega)K_{\lambda+2}(\omega) - K_{\lambda+1}(\omega)^2]/K_\lambda(\omega)^2$; see, e.g., Bibby and Sørensen (2003) and the references therein for further results and details on their derivation. For the purpose of this paper, we will standardize the generalized hyperbolic to have zero mean and unit variance and denote the standardized distribution as SGH. Its density is

$$f_{SGH}(x, \lambda, \omega, \rho) = f_{GHyp}(x, \lambda, \omega, \rho, \hat{\delta}, \hat{\mu}), \tag{4.8}$$

where $\hat{\delta} = [\omega^{-1}k_1(\omega) + \rho^2(1-\rho^2)^{-1}k_2(\omega)]^{-1/2}$ and $\hat{\mu} = -\rho(1-\rho^2)^{-1/2}\hat{\delta}k_1(\omega)$.

4.3.2 Specification of Factor Volatilities

As mentioned before, the GO-GARCH model does not restrict the dynamics of the univariate factors to follow simple GARCH(1,1) processes. Rather, the model is general enough to allow for the incorporation of arbitrary univariate dynamics. As such, we assume that $f_{it} = Z_{it}\sigma_{it}$, where the Z_{it} are independently distributed as SGH$(\lambda_i, \omega_i, \rho_i)$, and to capture the evolution of the scale parameters σ_{it}, we use the very successful Asymmetric Power ARCH, or A-PARCH,

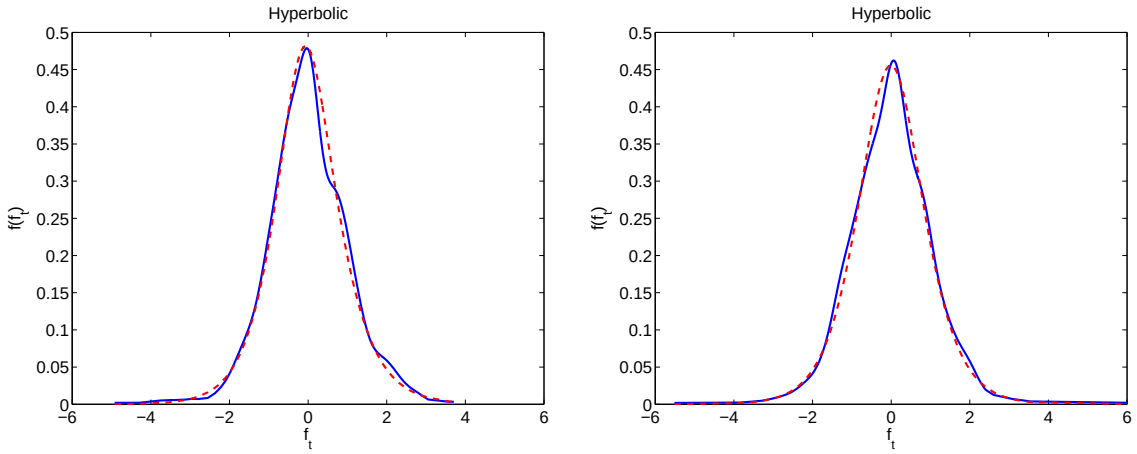

Figure 4.5: Kernel density (solid) of filtered residuals and fitted hyperbolic densities (dashes) for two ICs.

model proposed by Ding et al. (1993), given by

$$\sigma_{it}^{\delta_i} = c_{i0} + \sum_{j=1}^{r} c_{ij} \left(|f_{i,t-j}| - \gamma_{ij} f_{i,t-j} \right)^{\delta_i} + \sum_{j=1}^{s} d_{ij} \sigma_{i,t-j}^{\delta_i}$$

with $c_{ij} > 0$, $d_{ij} \geq 0$, $\delta_i > 0$, and $|\gamma_{ij}| < 1$.

4.4 Evaluating the GHyp Distribution and its Convolutions

4.4.1 Tail Probabilities

Calculating tail probabilities for $R_t = \mathbf{b}'\mathbf{r}_t$, the return on a portfolio of assets with weights $\mathbf{b} = [b_1, \ldots, b_d]'$, involves evaluating the cumulative distribution function (cdf) of a weighted sum of d independent random variables, each of which in this paper is assumed to be generalized hyperbolic. That is, we have $X_i \overset{\text{ind}}{\sim} \text{GHyp}(\lambda_i, \omega_i, \rho_i, \sigma_i, \mu_i)$, $i = 1, \ldots, d$, and are interested in the cdf of $S = \sum_{i=1}^{d} a_i X_i$. For $d = 1$, numeric integration of the density (4.6) can be used, but for $d > 1$, no expression exists for the convolution of such random variables when the distributional parameters in (4.6) are allowed to differ across assets—as required in practice.

There are three approaches to resolve this. The first is inverting the characteristic function of S, $\varphi_S(t)$, if it is available. From Gil-Pelaez (1951), the cdf of S is given by

$$F_S(x) = \frac{1}{2} + \frac{1}{2\pi} \int_0^\infty \frac{e^{itx}\varphi_S(-t) - e^{-itx}\varphi_S(t)}{it} \, \mathrm{d}t, \tag{4.9}$$

or, equivalently and computationally more efficient,

$$F_S(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty g(t) \, \mathrm{d}t, \tag{4.10}$$

where

$$g(t) = \frac{\operatorname{Im} z(t)}{t} \quad \text{and} \quad z = z(t) = e^{-itx}\varphi_S(t).$$

However, the integrand in equation (4.9) tends to be oscillatory, thus rendering numerical quadrature difficult. It is therefore expeditious to evaluate instead the following integral, due to Helstrom (1996):

$$F_S(x) = H(s_0) - \frac{1}{\pi}\int_0^\infty \operatorname{Re}\left[e^{I(s)}(1 - iCy)\right] \mathrm{d}y, \qquad s = s_0 + \frac{1}{2}Cy^2 + iy,$$

where $H(\cdot)$ denotes the Heaviside step function, and, with $\mathbb{K}_S$ denoting the cumulant generating function of S,

$$I(s) = \mathbb{K}_S(s) - xs - \ln(\operatorname{sgn}(s_0)s), \quad C = \frac{I^{(3)}(s_0)}{3I^{(2)}(s_0)},$$

superscripts in parentheses denoting derivatives. The saddlepoint s_0 lies in the convergence strip of the moment generating function and satisfies

$$\left.\frac{\mathrm{d}I(s)}{\mathrm{d}s}\right|_{s_0} = 0, \quad \operatorname{sgn}(s_0) = \operatorname{sgn}(X - \mathbb{E}[X]]).$$

The cumulant generating function, along with its first and second derivatives, is given in the Appendix. The third derivative is easily approximated numerically. The value of this method in the present context is in evaluating the accuracy of the (vastly faster) saddlepoint method described below.

The second method is via Fast Fourier Transfom, as used by Chen et al. (2006), a method which is time consuming if high accuracy is desired.

The third way is via the saddlepoint approximation, which is outlined in the Appendix. It can be thought of as approximate inversion of the characteristic function, but without requiring integration (thus being faster and avoiding potentially pathological integrand problems), or as an Edgeworth expansion, but vastly more accurate and without the problems associated with the latter, such as negative values of the density and poor accuracy in the tails. Its accuracy for $d > 1$ will be similar, if not higher, than for the $d = 1$ case, because, as assets are summed, a central limit effect takes place, drawing the distribution of S closer to normality—for which the SPA is *exact*. Thus, as a worst–case scenario, we present accuracy results for $d = 1$, for which we can compute the true values via (4.6) for the density and numeric integration of (4.6) for the cdf.

Figure 4.6 graphically illustrates the accuracy of the saddlepoint approximation. The left panel shows the true and approximate density, while the right panel shows the relative percentage error of the cdf approximation. We see that the cdf approximation has well under one percent relative error even far into the left tail, which is where accuracy is important in Value–at–Risk calculations.

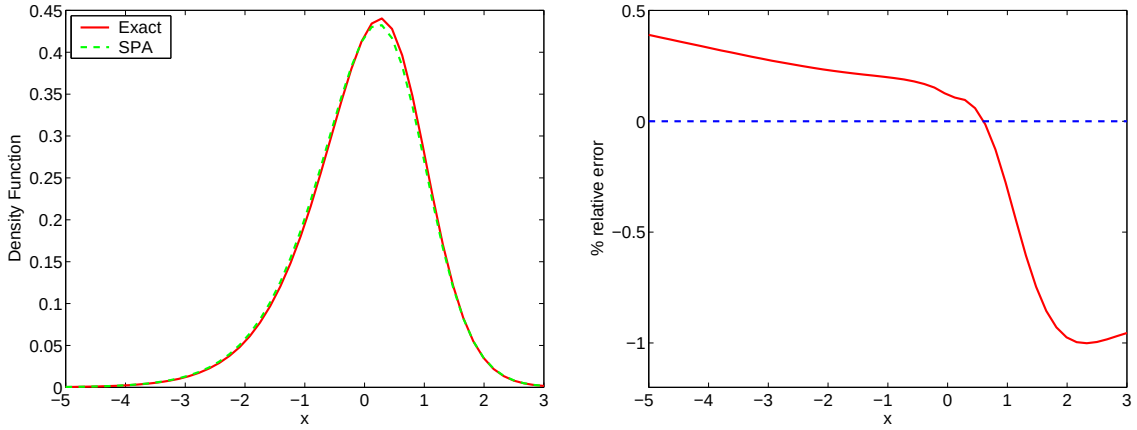

Figure 4.6: Left: The exact density (solid) and renormalized saddlepoint approximation (dashed) of the standardized GHyp density (4.8) for parameters $\lambda = 3$, $\omega = \sqrt{8}$ and $\rho = -1/3$. Right: The percentage relative error of the cdf saddlepoint approximation, defined as $100(\hat{F} - F)/\min(F, 1 - F)$, where $\hat{F}$ is given in (4.13) and the true cdf, F, was computed with numerical integration.

4.4.2 Quantiles and Expected Shortfall

In Value at Risk applications, the focus is on the quantiles of S, rather than on tail probabilities. Their computation involves numerically inverting the distribution function, i.e., solving

$$F_S(x_q) = q, \quad q \in (0, 1),$$

a computationally demanding task if exact methods are to be used for the evaluation of $F_S(\cdot)$. However, constructing a saddlepoint approximation for quantiles is no more involved than it is for tail areas, as detailed in the Appendix.

A related risk measure is the Expected Shortfall, defined, for a given VaR level q, as

$$\mathbb{E}\left[S|S \leq x_q\right] = \frac{1}{q}\int_{-\infty}^{x_q} S f_S(s)\mathrm{d}s, \tag{4.11}$$

which, in the case of the GHyp distribution, must be evaluated numerically, where f_S can be replaced by its saddlepoint approximation. However, unless renormalized, the saddlepoint approximation to the pdf does not integrate to unity, which may jeopardize the accuracy of such a calculation. It is therefore expedient to instead evaluate

$$\mathbb{E}\left[S|S \leq x_q\right] = x_q - \frac{1}{q}\int_{-\infty}^{x_q} F_S(s)\mathrm{d}s, \quad x_q < 0, \tag{4.12}$$

which follows from (4.11) upon integrating by parts. In (4.12), $F_S(s)$ can be safely replaced by its saddlepoint approximation, which is a proper distribution function. The third option is to use the direct saddlepoint approximation given in (4.17) below. Figure 4.7 plots the Expected Shortfall computed via equations (4.11), (4.12), and (4.17) (dotted, dash-dot, and solid lines,

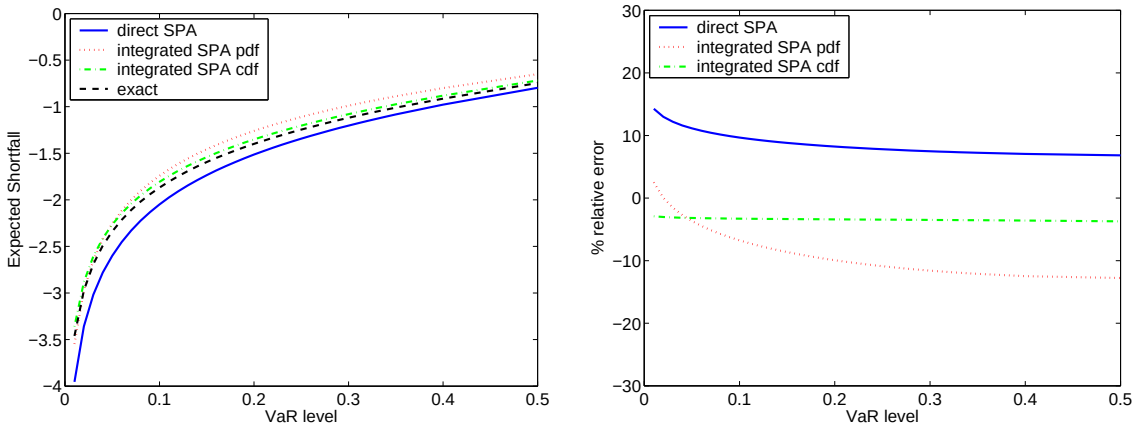

Figure 4.7: Approximations to Expected Shortfall (left panel) and percentage relative error (right panel) for a standardized NIG distribution with $\omega = 1.5$ and $\rho = -0.1$.

respectively) against the VaR level, for a standardized GHyp distribution with $\omega = 1.5$ and $\rho = -.1$. Approximation (4.12) clearly yields the highest accuracy and is therefore the recommended method.

4.5 Application

In order to exemplify the virtues of the CHICAGO method, we conduct a VaR backtesting exercise for a time series of Dow Jones 30 returns (3M, Alcoa, Altria, American Express, American International Group, AT&T, Boeing, Caterpillar, Citigroup, Coca-Cola, DuPont de Nemours, Exxon Mobil, General Electric, General Motors, Hewlett-Packard, Home Depot, Honeywell, Intel, IBM, Johnson & Johnson, J. P. Morgan Chase, McDonald's, Merck, Microsoft, Pfizer, Procter & Gamble, United Technologies, Verizon, Wal-Mart, Walt Disney), which we split alphabetically into 3 $d = 10$-dimensional samples. Each sample consists of daily returns for the period 9/23/92 to 3/23/07, resulting in a sample size of $T = 3,209$.

Univariate descriptive statistics for Samples 1, 2, and 3 are given in Tables 4.1, 4.3, and 4.5, respectively. Clearly, all series in the sample exhibit moderate skewness, and an amount of kurtosis incompatible with the assumption of Gaussianity. Consequently, the Jarque-Bera test rejects the null hypothesis of Gaussianity for all series; the critical value for this test at the 1% level is 9.21, and the smallest value observed in our samples is $1,060$. There is also strong evidence of conditional heteroskedasticity in the series, as is indicated by Ljung-Box Q tests (with 10 lags) on the squared residuals . The critical value of this test at the 1% level is 23.21, which is exceeded by all but two series in our sample. The correlation matrices for Samples 1, 2, and 3 are given in Tables 4.2, 4.4, and 4.6, respectively. Within all samples, the correlation between assets is fairly high, with average correlations ranging from 0.232 to 0.316.

We apply our CHICAGO method to each of the 3 $d = 10$-dimensional time series. Using a

	MMM	AA	MO	AXP	AIG	A	BA	CAT	C	KO
Mean	0.035	0.037	0.046	0.059	0.047	0.020	0.043	0.050	0.072	0.019
Std	1.537	2.136	1.944	2.007	1.744	1.855	1.999	2.023	2.063	1.558
Skewness	0.073	0.208	−0.312	−0.050	0.098	−0.098	−0.592	−0.229	0.043	−0.136
Kurtosis	6.942	5.878	11.346	6.791	6.771	6.431	11.161	6.865	8.890	7.772
JB Stat	2,076	1,128	9,349	1,919	1,902	1,575	9,077	2,021	4,631	3,049
LB Q_{10} Stat	138	260	205	1098	738	427	169	67	608	441

Table 4.1: Descriptive Statistics, Sample 1.

	MMM	AA	MO	AXP	AIG	A	BA	CAT	C	KO
MMM	1.000									
AA	0.339	1.000								
MO	0.203	0.145	1.000							
AXP	0.364	0.335	0.176	1.000						
AIG	0.365	0.304	0.216	0.541	1.000					
A	0.208	0.210	0.193	0.325	0.317	1.000				
BA	0.294	0.292	0.136	0.328	0.300	0.180	1.000			
CAT	0.380	0.423	0.173	0.362	0.334	0.197	0.303	1.000		
C	0.358	0.351	0.189	0.649	0.557	0.308	0.325	0.374	1.000	
KO	0.285	0.162	0.224	0.317	0.329	0.255	0.227	0.218	0.307	1.000

Table 4.2: Correlation Matrix, Sample 1. Average: 0.297, maximum: 0.649, minimum: 0.136.

moving estimation window of 1,000 observations, we compute 1-day ahead VaR forecasts for an equally weighted portfolio ($b_i = 1/10$) of the stocks in each sample. The results for the NIG and hyperbolic distributions are shown in the left and right columns of Figure 4.8, respectively, for Samples 1, 2, and 3, from top to bottom. VaR violations at the 1% and 5% levels are indicated by marks on the bottom and top axes, respectively. We will restrict our attention to the results for Sample 1, as those for the other samples are largely similar. At the nominal 1% (5%) level, the empirical levels of the VaR forecasts obtained from the CHICAGO method are 1.13% (4.48%) for the NIG, and 1.04% (3.98%) for the hyperbolic distribution, i.e., for this (and the other two) samples, the hyperbolic distribution yields slightly more conservative VaR forecasts. The Kupiec test accepts the null hypothesis of correct coverage with a p-value of 0.54 (0.26) for the NIG; for the hyperbolic model, the p-value is 0.85 (0.02), confirming that the NIG model has superior performance for this data set.

	DD	XOM	GE	GM	HPQ	HD	HON	INTC	IBM	JNJ
Mean	0.020	0.050	0.047	0.001	0.045	0.041	0.033	0.033	0.052	0.047
Std	1.771	1.473	1.702	2.108	2.657	2.181	2.154	2.154	1.982	1.473
Skewness	0.001	0.022	0.046	0.076	−0.086	−1.208	−0.301	−0.301	0.011	−0.425
Kurtosis	6.376	5.819	7.279	7.480	8.779	24.261	17.468	17.468	10.450	11.657
JB Stat	1,520	1,060	2,444	2,681	4,461	61,136	27,994	27,994	7,408	10,101
LB Q_{10} Stat	314	510	662	167	87	40	188	188	161	218

Table 4.3: Descriptive Statistics, Sample 2.

	DD	XOM	GE	GM	HPQ	HD	HON	INTC	IBM	JNJ
DD	1.000									
XOM	0.304	1.000								
GE	0.418	0.317	1.000							
GM	0.343	0.187	0.372	1.000						
HPQ	0.227	0.159	0.356	0.260	1.000					
HD	0.314	0.223	0.462	0.314	0.272	1.000				
HON	0.400	0.296	0.459	0.326	0.296	0.320	1.000			
INTC	0.400	0.296	0.459	0.326	0.296	0.320	1.000	1.000		
IBM	0.244	0.204	0.412	0.266	0.449	0.298	0.299	0.299	1.000	
JNJ	0.241	0.285	0.366	0.175	0.122	0.223	0.204	0.204	0.193	1.000

Table 4.4: Correlation Matrix, Sample 2. Average: 0.316, maximum: 0.462, minimum: 0.122.

	JPM	MCD	MRK	MSFT	PFE	PG	UTX	VZ	WMT	DIS
Mean	0.027	0.036	0.029	0.061	0.044	0.045	0.068	0.014	0.045	0.027
Std	1.122	1.727	1.815	2.164	1.872	1.660	1.809	1.767	1.896	2.025
Skewness	−0.421	−0.054	−1.697	−0.156	−0.222	−3.855	−1.997	0.088	0.023	−0.253
Kurtosis	7.721	7.819	33.278	8.855	6.182	88.813	40.607	7.240	7.036	11.497
JB Stat	3,068	3,101	123,953	4,589	1,377	991,292	190,976	2,402	2,174	9,672
LB Q_{10} Stat	695	198	4	240	304	16	96	421	340	152

Table 4.5: Descriptive Statistics, Sample 3.

	JPM	MCD	MRK	MSFT	PFE	PG	UTX	VZ	WMT	DIS
JPM	1.000									
MCD	0.134	1.000								
MRK	0.119	0.223	1.000							
MSFT	0.152	0.170	0.206	1.000						
PFE	0.136	0.226	0.526	0.254	1.000					
PG	0.029	0.269	0.300	0.130	0.311	1.000				
UTX	0.239	0.252	0.222	0.301	0.266	0.242	1.000			
VZ	0.117	0.185	0.212	0.238	0.229	0.195	0.238	1.000		
WMT	0.126	0.278	0.266	0.297	0.288	0.266	0.306	0.246	1.000	
DIS	0.233	0.227	0.201	0.305	0.237	0.153	0.345	0.255	0.285	1.000

Table 4.6: Correlation Matrix, Sample 3. Average: 0.232, maximum: 0.526, minimum: 0.029.

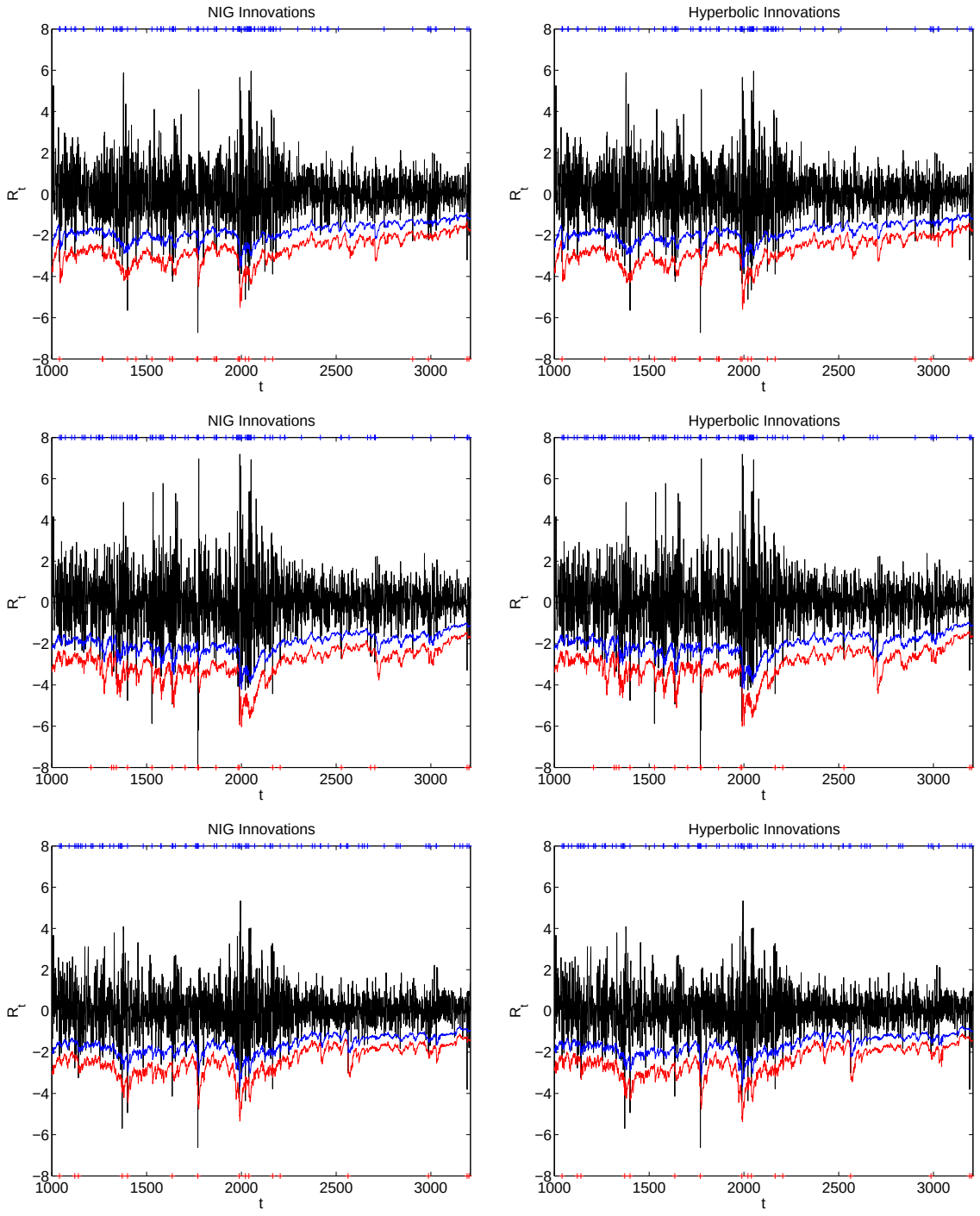

Figure 4.8: Returns, 1-day-ahead 1% and 5% VaR forecasts, and VaR violations, using NIG (left column) and hyperbolic (right column) innovations, for Samples 1, 2, and 3, from top to bottom. Empirical VaR (Kupiec p-value): **Sample 1**: 1.13% (0.54) and 4.48% (0.26) for NIG, 1.04% (0.85) and 3.98% (0.02) for hyperbolic. **Sample 2**: 0.95% (0.81) and 4.21% (0.08) for NIG, 0.81% (0.37) and 4.07% (0.04) for hyperbolic. **Sample 3**: 0.72% (0.17) and 4.16% (0.06) for NIG, 0.72% (0.17) and 4.21% (0.08) for hyperbolic.

4.6 Conclusions

The CHICAGO method developed in this paper is a fast, numerically reliable, and robust method for the estimation of Generalized Orthogonal GARCH models. The approach is completely modular: the estimation of the multivariate relationship between assets is entirely independent of the specification of the component dynamics and the distribution of innovations, thus not only simplifying model estimation, but also allowing empirical researchers to utilize whatever univariate volatility dynamics appear most adequate for the data at hand.

By virtue of the saddlepoint approximation to convolutions of independent generalized hyperbolic variates developed herein, accurate VaR forecasts for high-dimensional portfolios of assets can be obtained in a computationally efficient manner for different portfolio weights, thus allowing the procedure to be used in, e.g., real-time portfolio optimization.

A promising avenue for future research presents itself upon recognizing that the pattern of interdependence, and, hence, the mixing matrix in the GO-GARCH model, may not be stable over time. In this paper, we have addressed this possible shortcoming of the method by using a moving estimation window in our backtesting exercise. A more elaborate approach would consist in either incorporating a potential time dependence into the model itself, or in using a weighted estimator for the mixing matrix, thus giving more importance to recent observations.

Appendix 4.A Saddlepoint Approximations

If random variable X possesses a moment generating function, or, in short, mgf, given by $\mathbb{M}_X(t) = \mathbb{E}\big[\exp(tX)\big]$, then a highly accurate approximation to the density is given by

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi\,\mathbb{K}_X''(\hat{t})}} \exp\left\{\mathbb{K}_X(\hat{t}) - x\hat{t}\right\}, \quad x = \mathbb{K}_X'(\hat{t})\,. \tag{4.13}$$

This is referred to as the *(first order) saddlepoint density approximation* to f, abbreviated SPA, where $\hat{t} = \hat{t}(x)$ is the solution to the *saddlepoint equation* and is referred to as the *saddlepoint* at x. This method of approximation is attributed to Daniels (1954), though, via its similarity to the *Laplace method* of approximation, there is evidence that it can be traced back to Georg Bernhard Riemann (see the discussion and references in Kass, 1988, p. 235; and also Tierney, 1988). In general, $\hat{f}_X(x)$ will not integrate to one, although it will usually be close. It is, however, easy to renormalize it, by numeric integration.

The approximate cumulative distribution function (cdf) of X could be obtained by numerically integrating $\hat{f}$. However, in a celebrated paper, Lugannani and Rice (1980) derived a simple expression for the SPA to the cdf, given for continuous r.v.s by

$$\hat{F}_X(x) = \Pr(X < x) = \Phi(\hat{w}) + \phi(\hat{w})\left\{\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right\}, \qquad x \neq \mathbb{E}[X]\,, \tag{4.14}$$

where Φ and ϕ are the cdf and pdf of the standard normal distribution, respectively,

$$\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\hat{t}x - 2\mathbb{K}_X(\hat{t})} \quad \text{and} \quad \hat{u} = \hat{t}\sqrt{\mathbb{K}_X''(\hat{t})}.$$

In contrast to the density approximation (4.13), $\hat{F}_X(x)$ is such that, as x extends to the left and right edges of its support, $\hat{F}_X(x)$ tends to zero and one, respectively, so that renormalization is not necessary. Using above definition of $\hat{w}$, the density approximation (4.13) can also be written

$$\hat{f}_X(x) = \phi(\hat{w})/\sqrt{\mathbb{K}_X''(\hat{t})}. \tag{4.15}$$

An approximation to the q% quantile of X can be obtained by numerically solving

$$\hat{F}_X(x_q) = q, \quad q \in (0,1). \tag{4.16}$$

However, a direct application of (4.16) would result in a nested root search, because for every evaluation of the distribution function, the saddlepoint equation has to be solved. This can be circumvented by noting that the saddlepoint equation defines a bijection between $\hat{t}$ and x. Hence, $\hat{F}_X(x)$ can also be viewed as a function of $\hat{t}$, and it is equivalent to solve (4.16) in terms of $\hat{t}$, yielding, say, $\hat{t}_{x_q}$. The quantile approximation is then given by $x_q = \mathbb{K}_X'(\hat{t}_{x_q})$. The computational

advantages stem from the facts that firstly, $x = \mathbb{K}'_X(\hat{t})$ is in closed–form as a function of $\hat{t}$, and secondly, in many applications the convergence strip of the cgf is bounded, thus providing a bracketing interval for $\hat{t}_{x_q}$. Similarly, when integrating over the saddlepoint approximation, such as when calculating expected shortfall from (4.11) or (4.12), one may change variables and integrate over $\hat{t}$ instead of x.

Lastly, a direct saddlepoint approximation for Expected Shortfall is also available. Martin (2006) shows that

$$\mathbb{E}[X|X \leq c] \approx \mathbb{E}[X] - \frac{c - \mathbb{E}[X]}{\hat{t}} \frac{\hat{f}_X(c)}{\hat{F}_X(c)}. \tag{4.17}$$

Far more information on the SPA can be found in Reid (1988), Jensen (1995), Paolella (2007, Ch. 5) and Butler (2007).

Appendix 4.B Application to the GHyp Distribution

We will require derivatives of the Bessel function. It is straightforward to verify that

$$-2K'_\nu(x) = K_{\nu-1}(x) + K_{\nu+1}(x), \quad \nu \in \mathbb{R},\ x \in \mathbb{R}_{>0}. \tag{4.18}$$

First consider the $d = 1$ case. Let $X \sim \text{GHyp}(\lambda, \omega, \rho, \sigma, \mu)$ with density (4.6). With $\beta \equiv \omega\delta^{-1}\rho(1-\rho^2)^{-1/2}$ and $\psi \equiv \omega^2\delta^{-2}$, the moment generating function of X is given by

$$\mathbb{M}_X(t) = e^{\mu t} \frac{K_\lambda\left(\omega\sqrt{1 - \frac{2\beta t + t^2}{\psi}}\right)}{K_\lambda(\omega)\left(1 - \frac{2\beta t + t^2}{\psi}\right)^{\lambda/2}}, \tag{4.19}$$

with convergence strip given by those values of t such that

$$1 - \frac{2\beta t + t^2}{\psi} > 0 \overset{\psi>0}{\Rightarrow} t^2 + 2\beta t - \psi < 0.$$

The solutions of $t^2 + 2\beta t - \psi = 0$ are $t = -\beta \pm \sqrt{\beta^2 + \psi}$, so that the convergence strip is

$$-\beta - \sqrt{\beta^2 + \psi} < t < -\beta + \sqrt{\beta^2 + \psi}. \tag{4.20}$$

The cumulant generating function, or cgf, is defined as $\mathbb{K}_X(t) = \ln \mathbb{M}_X(t)$. The cgf corresponding to (4.19) is

$$\mathbb{K}_X(t) = \mu t + \ln K_\lambda(\omega Q) - \ln K_\lambda(\omega) - \lambda \ln(Q), \tag{4.21}$$

where $Q = Q(t) := \sqrt{1 - (2\beta t + t^2)/\psi}$. It is easy to see that

$$\frac{dQ(t)}{dt} = \frac{1}{2}\left(1 - \frac{2\beta t + t^2}{\psi}\right)^{-1/2}\left(-\frac{2\beta + 2t}{\psi}\right) = -\frac{\beta + t}{Q\psi},$$

so, via (4.18) and some simplification,

$$\mathbb{K}'_X(t) = \mu + \frac{\beta+t}{Q\psi}\left(\frac{\omega}{2}\frac{K_{\lambda-1}(\omega Q)+K_{\lambda+1}(\omega Q)}{K_\lambda(\omega Q)}+\frac{\lambda}{Q}\right).$$

Numerically solving $\mathbb{K}'_X(t) = x$ in the range (4.20) then gives the saddlepoint $\hat{t}$. It can be proven that there exists one, and only one, solution to the saddlepoint equation for t restricted to the convergence strip of the mgf, so that this root search is well–defined and computationally straightforward.

Tedious but straightforward algebra then shows that

$$\mathbb{K}''_X(t) = \frac{\beta+t}{Q\psi}\times P_1 + \left[\frac{\omega}{2}\frac{K_{\lambda-1}(\omega Q)+K_{\lambda+1}(\omega Q)}{K_\lambda(\omega Q)}+\frac{\lambda}{Q}\right]\times\left[\frac{1}{Q\psi}\left(1+\frac{(\beta+t)^2}{Q^2\psi}\right)\right],$$

where

$$P_1 = \frac{\omega}{2}P_2 + \frac{\lambda(\beta+t)}{Q^3\psi}$$

and

$$\begin{aligned} K_\lambda^2(\omega Q)P_2 &= K_\lambda(\omega Q)\times\left[\frac{\omega}{2}\left(\frac{\beta+t}{Q\psi}\right)\left(K_{\lambda-2}(\omega Q)+2K_\lambda(\omega Q)+K_{\lambda+2}(\omega Q)\right)\right] \\ &\quad -\left[K_{\lambda-1}(\omega Q)+K_{\lambda+1}(\omega Q)\right]\times\left[\frac{\omega}{2}\left(K_{\lambda-1}(\omega Q)+K_{\lambda+1}(\omega Q)\right)\left(\frac{\beta+t}{Q\psi}\right)\right]. \end{aligned}$$

With these expressions, the first–order SPA to the pdf and cdf can be computed from (4.13) and (4.14), respectively.

The application to the $d > 1$ case is now straightforward: Because of independence, the cgf of $S = \sum_{i=1}^d a_i X_i$, $a_i \neq 0$, is just $\mathbb{K}_S(t) = \sum_{i=1}^d \mathbb{K}_{X_i}(a_i t)$, where the cgf of each X_i is given in (4.21). The singularities of the cgf lie on both sides of the origin at the points $\left(-\beta_i \pm \sqrt{\beta_i^2+\psi_i}\right)/a_i$.

In the special case of the NIG distribution, matters simplify considerably. Specifically,

$$\mathbb{K}_X(t) = \mu t - \omega Q + \omega, \quad \mathbb{K}'_X(t) = \mu + \omega\frac{\beta+t}{Q\psi}, \quad \mathbb{K}''_X = \frac{\omega}{Q\psi}+\frac{\omega(\beta+t)^2}{Q^3\psi^2},$$

and the saddlepoint is now given in closed form as

$$\hat{t} = z\frac{\bar{\alpha}}{\bar{y}\delta} - \beta,$$

where z, $\bar{y}$ and $\bar{\alpha}$ are as in (4.6). The other quantities entering and (4.14) and (4.15) simplify to

$$\mathbb{K}''_X(\hat{t}) = \bar{y}^3\delta^2\bar{\alpha} \quad \text{and} \quad \hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2(\bar{y}\bar{\alpha} - z\rho\bar{\alpha} - \omega)}.$$

For completeness, we also give the simplified expressions for the quantities $\hat{\kappa}_i \equiv \mathbb{K}_X^{(i)}(\hat{t})\mathbb{K}''_X(\hat{t})^{-i/2}$,

$i \in \{3, 4\}$, required for evaluating the second-order equivalents of (4.13) and (4.14):

$$\hat{\kappa}_3 = 3\frac{z}{\sqrt{\bar{y}\bar{\alpha}}} \quad \text{and} \quad \hat{\kappa}_4 = 3\frac{1+5z^2}{\bar{y}\bar{\alpha}}.$$

Bibliography

Aas, K., I. Haff, and X. K. Dimakos (2006): "Risk Estimation using the Multivariate Normal Inverse Gaussian Distribution," *Journal of Risk*, 8. 84

Adler, R. J., R. E. Feldman, and M. S. Taqqu, eds. (1998): *A Practical Guide to Heavy Tails*, Boston: Birkhäuser. 19

Ahn, S. C. and P. Schmidt (1995): "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 5–27. 28

Anderson, T. W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, 76, 598–606. 28

Andrews, D. W. K. (1993): "Exactly Median-Unbiased Estimation of First Order Autoregressive / Unit Root Models," *Econometrica*, 61, 139–65. 2, 8, 9, 10, 11, 13, 21, 22, 23, 29, 31, 33, 55

Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press. 28

Arellano, M. and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277–297. 28, 42

Bao, Y., T.-H. Lee, and B. Saltoglu (2003): "A Test for Density Forecast Comparison with Applications to Risk Management," Working paper, University of California, Riverside. 73

Barndorff-Nielsen, O. E. (1986): "Inference on Full and Partial Parameters Based on the Standardized Signed Log Likelihood Ratio," *Biometrika*, 73, 307–322. 49

——— (1990): "Approximate Interval Probabilities," *Journal of the Royal Statistical Society B*, 52, 485–496. 49

——— (1998): "Processes of Normal Inverse Gaussian Type," *Finance & Stochastics*, 2, 41–68. 84

Barndorff-Nielsen, O. E. and D. R. Cox (1989): *Asymptotic Techniques for Use in Statistics*, London: Chapman & Hall. 11

Bauer, C. (2000): "Value at Risk Using Hyperbolic Distributions," *Journal of Economics and Business*, 52, 455–467. 84

Bauwens, L., S. Laurent, and J. V. K. Rombouts (2006): "Multivariate GARCH Models: A Survey," *Journal of Applied Econometrics*, 21, 79–109. 76

Bhargava, A., L. Franzini, and W. Narendranathan (1982): "Serial Correlation and the Fixed Effects Model," *The Review of Economic Studies*, 49, 533–549. 53

Bibby, B. M. and M. Sørensen (2003): "Hyperbolic Processes in Finance," in *Handbook of Heavy Tailed Distributions in Finance*, ed. by S. T. Rachev, Amsterdam: Elsevier Science. 85

Bollerslev, T. (1990): "Modelling the Coherence in Short-Run Nominal Exchange Rate: A Multivariate Generalized ARCH Approach," *Review of Economics and Statistics*, 72, 498–505. 76

Bollerslev, T., R. F. Engle, and J. Wooldridge (1988): "A Capital Asset-Pricing Model with Time-Varying Covariances," *Journal of the Political Economy*, 96, 116–131. 76

Boswijk, H. P. and R. van der Weide (2006): "Wake me up before you GO-GARCH," Tinbergen Institute Discussion Paper TI 2006-79/4. 79, 80, 83

Breitung, J. (2002): "Nonparametric Tests for Unit Roots and Cointegration," *Journal of Econometrics*, 108, 343–63. 8

Bruno, G. S. F. (2005): "Approximating the Bias of the LSDV Estimator for Dynamic Unbalanced Panel Data Models," *Economics Letters*, 87, 361–366. 35

Bun, M. J. G. and M. A. Carree (2005): "Bias-Corrected Estimation in Dynamic Panel Data Models," *Journal of Business and Economic Statistics*, 23, 200–210. 28, 35, 42

Bun, M. J. G. and J. F. Kiviet (2003): "On the Diminishing Returns of Higher-Order Terms in Asymptotic Expansions of Bias," *Economics Letters*, 79, 145–152. 35

Butler, R. W. (2007): *Saddlepoint Approximations with Applications*, Cambridge: Cambridge University Press. 95

Butler, R. W. and M. S. Paolella (2002): "Calculating the Density and Distribution Function for the Singly and Doubly Noncentral F," *Statistics and Computing*, 12, 9–16. 74

Butler, R. W. and A. T. A. Wood (2002): "Laplace Approximation for Hypergeometric Functions of Matrix Argument," *Annals of Statistics*, 30, 1155–1177. 67, 68, 72

Cappuccio, N., D. Lubian, and D. Raggi (2004): "Asymmetry and Heavy Tails in Financial Returns: Evidence from International Stock Markets," Working paper, Department of Economics, University of Padova. 73

Chen, Y., W. Härdle, and V. Spokoiny (2006): "GHICA — Risk Analysis with GH Distributions and Independent Components," SFB 649 Discussion Paper 2006-78, Humboldt-University, Berlin. 76, 87

Choi, C.-Y., N. C. Mark, and D. Sul (2004): "Bias Reduction by Recursive Mean Adjustment in Dynamic Panel Data Models," *Economics Working Paper Archive at WUSTL*. 28

Daniels, H. E. (1954): "Saddlepoint Approximation in Statistics," *Ann. Math. Statist.*, 25, 631–650. 1, 65, 69, 94

——— (1987): "Tail Probability Approximation," *International Statistical Review*, 55, 37–48. 24

Daniels, H. E. and G. A. Young (1991): "Saddlepoint Approximation for the Studentized Mean, with an Application to the Bootstrap," *Biometrika*, 78, 169–179. 61

Davis, R. and S. Resnick (1986): "Limit Theory for the Sample Covariance and Correlation Functions of Moving Averages," *The Annals of Statistics*, 14, 533–558. 18

DiCiccio, T. J. and M. A. Martin (1991): "Approximations of Marginal Tail Probabilities for a Class of Smooth Functions with Applications to Bayesian and Conditional Inference," *Biometrika*, 78, 891–902. 63

Ding, Z., C. W. Granger, and R. F. Engle (1993): "A Long Memory Property of Stock Market Returns and a New Model," *Journal of Empirical Finance*, 1, 83–106. 71, 77, 86

Dowd, K. (2005): *Measuring Market Risk*, New York: Wiley, 2nd ed. 71

Dubbelman, C., A. S. Louter, and A. P. J. Abrahamse (1978): "On Typical Characteristics of Economic Time Series and the Relative Qualities of Five Autocorrelation Tests," *Journal of Econometrics*, 8, 295–306. 19

Dufour, J.-M. and M. L. King (1991): "Optimal Invariant Tests for the Autocorrelation Coefficient in Linear Regressions with Stationary or Nonstationary AR(1) Errors," *Journal of Econometrics*, 47, 115–143. 22

Durbin, J. and G. S. Watson (1971): "Testing for Serial Correlation in Least Squares Regression. III," *Biometrika*, 58, 1–19. 19

Elliot, G. and J. H. Stock (2001): "Confidence Interval for Autoregressive Coefficients Near One," *Journal of Econometrics*, 103, 155–81. 8

Engle, R. and F. K. Kroner (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150. 76, 77

Engle, R. F. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics*, 20, 339–350. 76

Engle, R. F., V. K. Ng, and M. Rothschild (1990): "Asset Pricing with a factor-arch covariance structure: Empirical estimates for treasury bills," *Journal of Econometrics*, 45, 213–237. 76, 77

Everaert, G. and L. Pozzi (2007): "Bootstrap-Based Bias Correction for Dynamic Panels," *Journal of Economic Dynamics and Control*, 31, 1160–1184. 28

Gil-Pelaez, J. (1951): "Note on the Inversion Theorem," *Biometrika*, 38, 481–82. 86

Glasserman, P. (2004): "Tail Approximations for Portfolio Credit Risk," *Journal of Derivatives*, 12, 24–42. 1

Gordy, M. B. (2002): "Saddlepoint Approximation of CreditRisk+," *Journal of Banking and Finance*, 26, 1335–1353. 1

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118. 31

Gouriéroux, C., P. C. B. Phillips, and J. Yu (2006): "Indirect Inference for Dynamic Panel Models," Cowles Foundation Discussion Paper No. 1550. 28, 34

Goutis, C. and G. Casella (1999): "Explaining the Saddlepoint Approximation," *The American Statistician*, 53, 216–224. 11

Granger, C. W. J. and Z. Ding (1995): "Some Properties of Absolute Return, An Alternative Measure of Risk," *Annales D'économie et de Statistique*, 40, 67–91. 84

Haas, M., S. Mittnik, and M. S. Paolella (2006): "Modeling and Predicting Market Risk With Laplace-Gaussian Mixture Distributions," *Applied Financial Economics*, 16, 1145–1162. 84

Hahn, J. and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when both n and T are large," *Econometrica*, 70, 1639–1657. 28, 35, 42

Hansen, B. E. (1999): "The Grid Bootstrap and the Autoregressive Model," *The Review of Economics and Statistics*, 81, 594–607. 8

Harvey, C. R. and A. Siddique (1999): "Autoregressive Conditional Skewness," *Journal of Financial and Quantitative Analysis*, 34, 465–487. 71

Helstrom, C. W. (1996): "Calculating the Distribution of the Serial Correlation Estimator by Saddlepoint Integration," *Econometric Theory*, 12, 458–80. 87

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscoglu (2002): "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods," *Journal of Econometrics*, 109, 107–150. 28

Hyvärinen, A. (1999): "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, 10, 626–634. 80

Hyvärinen, A., J. Karhunen, and E. Oja (2001): *Independent Component Analysis*, New York: John Wiley & Sons. 79, 80

Imhof, J. P. (1961): "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–26. 8, 11, 13, 33

Jensen, J. L. (1992): "The Modified Signed Likelihood Statistic and Saddlepoint Approximations," *Biometrika*, 79, 693–703. 49

——— (1995): *Saddlepoint Approximations*, Oxford: Oxford University Press. 11, 95

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995): *Continuous Univariate Distributions*, vol. 2, Wiley, 2nd ed. 67

Kass, R. E. (1988): "Comment on: Saddlepoint Methods and Statistical Inference," *Statistical Science*, 3, 234–236. 94

Kim, T.-H., S. Leybourne, and P. Newbold (2002): "Unit Root Tests with a Break in Innovation Variance," *Journal of Econometrics*, 109, 365–387. 8

King, M. L. (1985): "A Point Optimal Test for Autoregressive Disturbances," *Journal of Econometrics*, 27, 21–37. 19

Kiviet, J. F. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53–78. 28, 35

Kiviet, J. F. and G. D. A. Phillips (1990): "Exact Similar Tests for the Root of a First–Order Autoregressive Regression Model," Unpublished manuscript, Department of Actuarial Science and Econometrics, University of Amsterdam. 10

Kocherlakota, K. and S. Kocherlakota (1991): "On the Doubly Noncentral t distribution," *Communications in Statistics: Simulation and Computation*, 20, 23–31. 60, 67

Kotz, S., K. Podgorski, and T. Kozubowski (2001): *The Laplace Distribution and Generalizations: A Revisit with Application to Communication, Economics, Engineering and Finance*, Boston: Birkhäuser. 19

Krishnan, M. (1967): "The Moments of a Doubly Noncentral t-Distribution," *Journal of the American Statistical Association*, 62, 278–287. 61, 66, 67

——— (1968): "Series Representations of the Doubly Noncentral t-Distribution," *Journal of the American Statistical Association*, 63, 1004–1012. 61

Kruiniger, H. (2002): "On the Estimation of Panel Regression Models with Fixed Effects," *Working Paper, Queen Mary, University of London*. 28

Küchler, U., K. Neumann, M. Sørensen, and A. Streller (1999): "Stock Returns and Hyperbolic Distributions," *Mathematical and Computer Modelling*, 29, 1–15. 84

Kuester, K., S. Mittnik, and M. S. Paolella (2006): "Value–at–Risk Prediction: A Comparison of Alternative Strategies," *Journal of Financial Econometrics*, 4, 53–89. 71

Kurozumi, E. (2002): "Testing for Stationarity with a Break," *Journal of Econometrics*, 108, 63–99. 8

Ledoit, O. and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection." *Journal of Empirical Finance*, 10, 603–621. 78

Lieberman, O. (1994a): "Saddlepoint Approximation for the Distribution of a Ratio of Quadratic Forms in Normal Variables," *Journal of the American Statistical Association*, 89, 924–928. 34, 46

——— (1994b): "Saddlepoint Approximation for the Least Squares Estimator in First-Order Autoregression," *Biometrika*, 81, 807–11. 11, 14, 24

Lisi, F. (2005): "Testing Asymmetry in Financial Time Series," Working paper, Department of Statistical Sciences, University of Padua. 73

Lugannani, R. and S. O. Rice (1980): "Saddlepoint Approximations for the Distribution of Sums of Independent Random Variables," *Adv. Appl. Prob.*, 12, 475–90. 24, 65, 94

MacKinnon, J. G. and A. A. Smith, Jr. (1998): "Approximate Bias Correction in Econometrics," *Journal of Econometrics*, 85, 205–30. 13, 31

Martin, R. (2006): "The Saddlepoint Method and Portfolio Optionalities," *Risk Magazine*, 19, 93–95. 95

McCullagh, P. and R. J. Tibshirani (1990): "A Simple Method for the Adjustment of Profile Likelihoods," *Journal of the Royal Statistical Society Series B*, 52, 325–44. 29, 31, 40, 42

McDonald, J. B. (1997): "Probability Distributions for Financial Models," in *Handbook of Statistics*, ed. by G. Maddala and C. Rao, Elsevier Science, vol. 14. 19

Mittnik, S. and M. S. Paolella (2000): "Conditional Density and Value-at-Risk Prediction of Asian Currency Exchange Rates," *Journal of Forecasting*, 19, 313–333. 73

Mittnik, S., M. S. Paolella, and S. T. Rachev (1998): "Unconditional and Conditional Distributional Models for the Nikkei Index," *Asia-Pacific Financial Markets*, 5, 99–128. 71, 84

Moon, H. R. and P. C. B. Phillips (2004): "GMM Estimation of Autoregressive Roots Near Unity with Panel Data," *Econometrica*, 72, 467–522. 28

Morgan, B. J. T. (2000): *Applied Stochastic Modelling*, London: Arnold. 72

Mudholkar, G. S. and Y. P. Chaubey (1976): "A Simple Approximation for the Doubly Noncentral-t Distribution," *Communications in Statistics: Simulation and Computation*, 5, 85–92. 61, 68

Nandi, S. B. and S. Choudhury (2002): "Series Representation of Doubly Noncentral t Distribution by Use of the Mellin Integral Transform," *Communications in Statistics: Theory and Methods*, 31, 2139–2151. 61

Neyman, J. and E. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32. 28

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426. 28, 35

Paige, R. and A. Trindade (2006): "Practical Small-Sample Inference for Order One Subset Autoregressive Models via Saddlepoint Approximations," Working Paper. 29

Palm, F. C. (1996): "GARCH Models of Volatility," in *Handbook of Statistics*, ed. by G. Maddala and C. Rao, Amsterdam: Elsevier Science, vol. 14, 209–240. 71, 84

Paolella, M. (2003): "Computing Moments of Ratios of Quadratic Forms in Normal Variables," *Computational Statistics and Data Analysis*, 42, 313–331. 13

Paolella, M. S. (2007): *Intermediate Probability: A Computational Approach*, New York: John Wiley & Sons. 95

Patnaik, P. B. (1949): "The Noncentral χ^2 and F-Distributions and Their Applications," *Biometrika*, 36, 202–232. 68

Perera, D., M. Peiris, J. Robinson, and N. Weber (2006): "Saddlepoint Approximation Methods for Testing of Serial Correlation in Panel Time Series Data," *Journal of Statistical Computation and Simulation*, 76, 1001–1013. 29

Phillips, P. C. B. and D. Sul (2003): "Dynamic Panel Estimation and Homogeneity Testing Under Cross Section Dependence," *Econometrics Journal*, 6, 217–259. 28, 31, 33, 35

——— (2007): "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross Section Dependence," *Journal of Econometrics*, 137, 162–188. 35

Pitman, E. J. G. (1937): "The "Closest" Estimates of Statistical Parameters," *Proc. Camb. Phil. Soc.*, 33, 212–222. 39

Praetz, P. D. (1972): "The Distribution of Share Price Changes," *The Journal of Business*, 45, 49–55. 73

Prause, K. (1999): "The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures," Ph.D. thesis, Albert-Ludwigs-Universität Freiburg. 84

Premaratne, G. and A. Bera (2005): "A Test for Symmetry with Leptokurtic Financial Data," *Journal of Financial Econometrics*, 3, 169–187. 73

Reid, N. (1988): "Saddlepoint Methods and Statistical Inference (with Discussion)," *Statistical Science*, 3, 213–238. 95

Reimann, S. (2005): "Evidence for a Hyperbolic-like Distribution of Asset Returns Drawn from a Simple Economical Financial Markets Model," Institut for Empirical Research in Economics Working Paper No. 232, University of Zurich, Switzerland. 84

Robbins, M. (1948): "The Distribution of Student's t When the Population Means are Unequal," *Annals of Mathematical Statistics*, 19, 406–410. 60

Roberts, L. A. (1995): "On the Existence of Moments of Ratios of Quadratic Forms," *Econometric Theory*, 11, 750–774. 18

Routledge, R. and M. Tsao (1997): "On the Relationship Between two Asymptotic Expansions for the Distribution of Sample Mean and its Applications," *Annals of Statistics*, 25, 2200–2209. 65, 69

Sadefo-Kamdem, J. (2006): "VaR and ES for Linear Portfolios with Mixture of Elliptic Distributions Risk Factors," *European Journal of Operational Research*, forthcoming. 84

Saikkonen, P. and H. Lütkepohl (2002): "Testing for a Unit Root in a Time Series with a Level Shift at Unknown Time," *Econometric Theory*, 18, 313–48. 8

Sawa, T. (1978): "The Exact Moments of the Least Squares Estimator for the Autoregressive Model," *Journal of Econometrics*, 8, 159–72. 13, 34

Scheffé, H. (1959): *The Analysis of Variance*, New York: John Wiley & Sons, Inc. 60

Schmidt, R., T. Hrycej, and E. Stützle (2006): "Multivariate Distribution Models with Generalized Hyperbolic Margins," *Computational Statistics & Data Analysis*, 50, 2065–2096. 84

Shively, P. A. (2001): "Trend-Stationary GNP: Evidence from a new exact Pointwise Most Powerful Invariant Unit Root Test," *Journal of Applied Econometrics*, 16, 537–51. 8

Shyam-Sunder, L. and S. C. Myers (1999): "Testing Static Tradeoff Against Pecking Order Models of Capital Structure," *Journal of Financial Economics*, 51, 219–244. 1

Straumann, D. (2005): *Estimation in Conditionally Heteroscedastik Time Series Models*, Berlin, Heidelberg: Springer. 73

Strawderman, R. L. and M. T. Wells (1998): "Approximately Exact Inference for the Common Odds Ratio in Several 2×2 Tables," *Journal of the American Statistical Association*, 93, 1294–1307. 29

Tanizaki, H. (2000): "Bias Correction of OLSE in the Regression Model with Lagged Dependent Variables," *Journal of Computational Statistics and Data Analysis*, 34, 495–511. 8, 13, 34

Temme, N. M. (1982): "The Uniform Asymptotic Expansion of a Class of Integrals Related to Cumulative Distribution Functions," *SIAM Journal on Mathematical Analysis*, 13, 239–253. 49

Tierney, L. (1988): "Comment on: Saddlepoint Methods and Statistical Inference," *Statistical Science*, 3, 233–234. 94

Tierney, L., R. E. Kass, and J. B. Kadane (1989a): "Approximate Marginal Densities of Nonlinear Functions," *Biometrika*, 76, 425–433. 47

——— (1989b): "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716. 53

Ullah, A., V. K. Srivastava, and N. Roy (1995): "Moments of the Function of Non-Normal Random Vector with Applications to Econometric Estimators and Test Statistics," *Econometric Reviews*, 14, 459–471. 18

van der Weide, R. (2002): "GO-GARCH: A Multivariate Generalized Orthogonal GARCH Model," *Journal of Applied Econometrics*, 17, 549–564. 76, 77, 78

Vrontos, I. D., P. Dellaportas, and D. N. Politis (2003): "A Full–Factor Multivariate GARCH model," *The Econometrics Journal*, 6, 312–334. 77

Wilson, G. T. (1989): "On the use of Marginal Likelihood in Time Series Model Estimation," *Journal of the Royal Statistical Society B*, 51, 15–27. 40, 42

Curriculum Vitae

Simon Broda, born January 21st, 1977 in Kiel, Germany

Education

2003–present	Doctoral student in Finance, University of Zürich, Switzerland
2002–2003	Doctoral student in Finance, University of Kiel, Germany
1997–2002	University of Kiel, Germany (Diplom-Kaufmann)
1987–1996	Gymnasium Kiel-Wellingdorf, Germany (Allgemeine Hochschulreife)

Academic Positions and Other Work Experience

2003–present	Research assistant, Swiss Banking Institute, University of Zürich, Switzerland
2002–2003	Research assistant, Chair for Finance, University of Kiel, Germany
1999, 2000	Intern, PricewaterhouseCoopers, Hamburg, Germany